Project 1

Anna Leisa Sauser

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

Develop a set of Logistic regression model from the 89 specimens that you can use to predict the group membership of the remaining 199 specimens'.

1. Format the data contained in the excel spreadsheet for use in R.
2. Perform an exploratory data analysis.

   To format the data, I read in the training and testing data sets. My first step was to look at the data and get a visual sense for what the data looked like. I saw some patterns in the data (e.g., values for a given measurement seem to fall within a particular, noticeable range), and I took note of some missing values. I also did some internet searching to understand how the various measurements were taken and to read some additional background on voles.

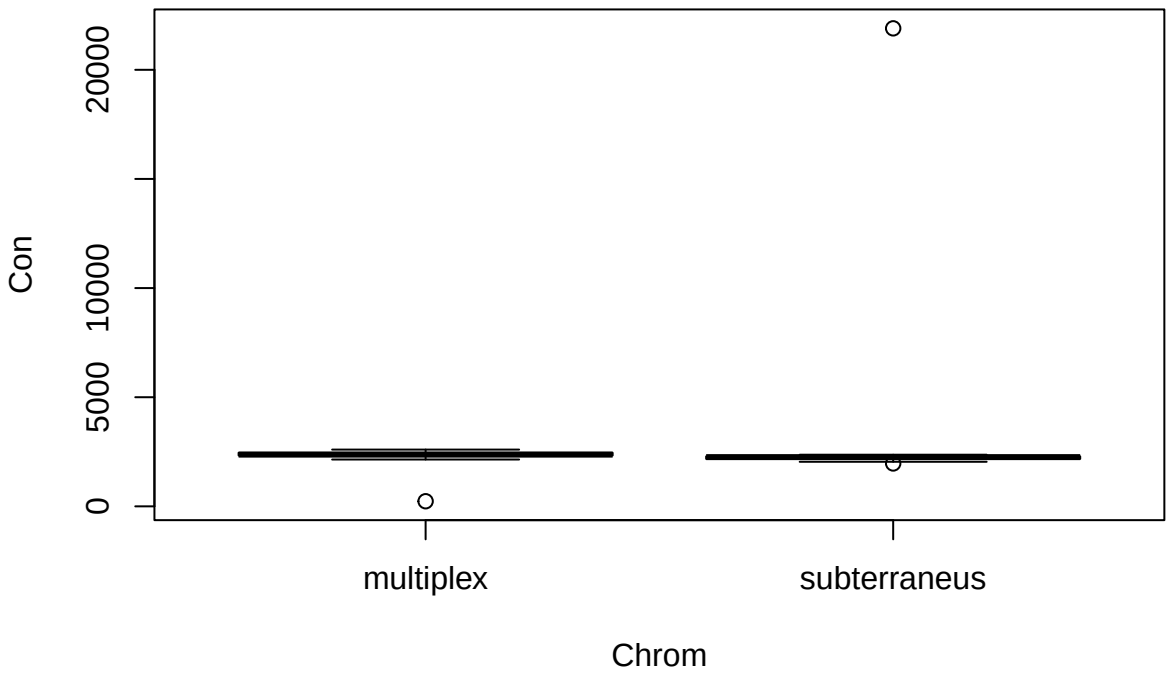
   I then contemplated how best and most accurately to clean the data. I considered using a mean value to replace N/A values, but decided to omit N/A values and work with complete lines of data.

   My process for reading the data in and beginning to clean the data:

```
library(readxl)
setwd("~/Desktop")
test_data <- read_excel("Vole Skulls.xlsm")
training_data1 <- read_excel("Vole Skulls.xlsm", sheet = "Subterraneus")
training_data1 <- na.omit(training_data1)
training_data2 <- read_excel("Vole Skulls.xlsm", sheet = "Multiplex")
training_data2 <- na.omit(training_data2)
colnames(training_data2) <- colnames(training_data1) # Change column names
training_data <- rbind(training_data1, training_data2)
colnames(test_data) <- c("Index", "Chrom", "Con", "Height", "Width")
colnames(training_data) <- c("Index", "Chrom", "Con", "Height", "Width")
```

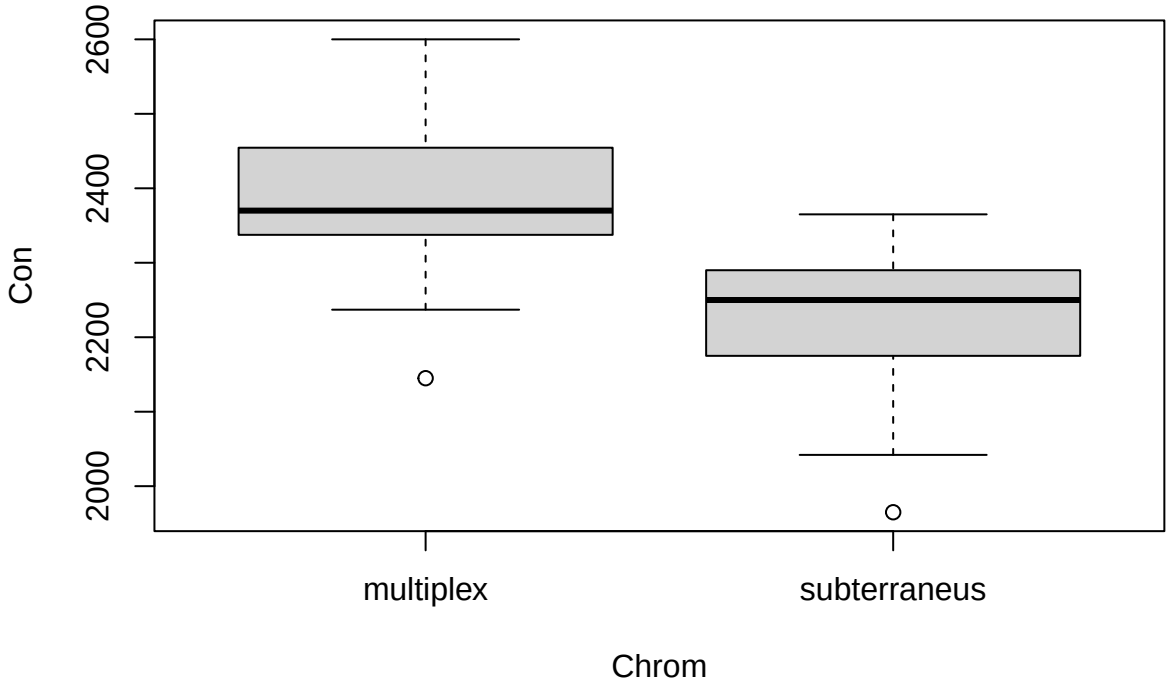
After omitting the N/A values, I created a boxplot for each of the different measurements: Condylo incisive length ("Con" in my analysis), Skull Height ("Height") and Skull Width ("Width"). Each boxplot allowed me to see the patterns in the data more clearly, and to see where there were outliers. Because it was a small dataset, I looked at the dataset to see where these outliers were and removed them.

```
boxplot(Con ~ Chrom, data = training_data)
```



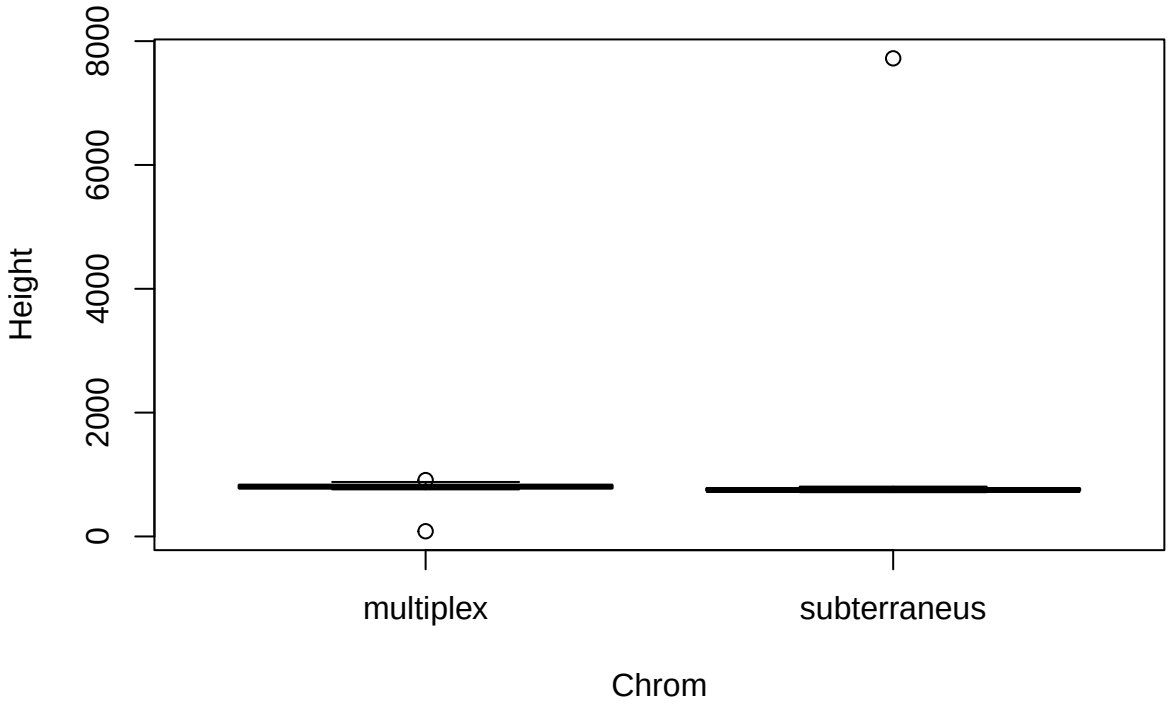
Here, I removed rows 13 and 86 as an initial step to clean up the data. You can see the resulting boxplots show clean groups of data.

```
data1 <- training_data[-c(13, 86),]
boxplot(Con ~ Chrom, data = data1)
```

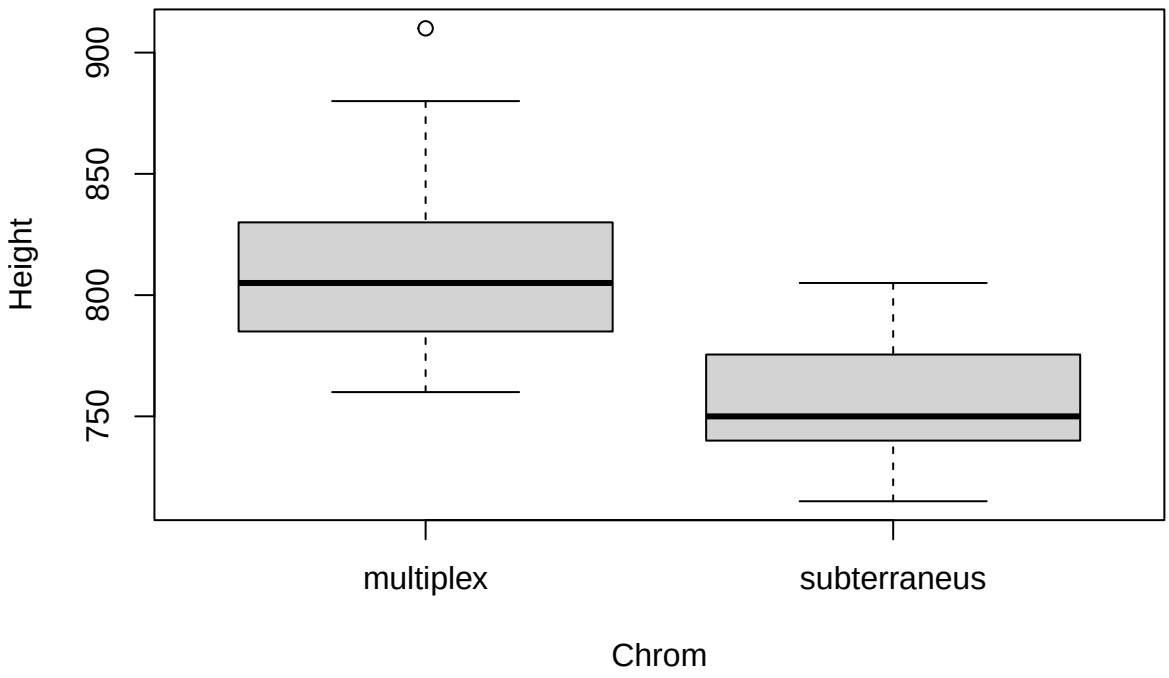


I followed those steps with Height and Width, and again, removed outlying data manually.

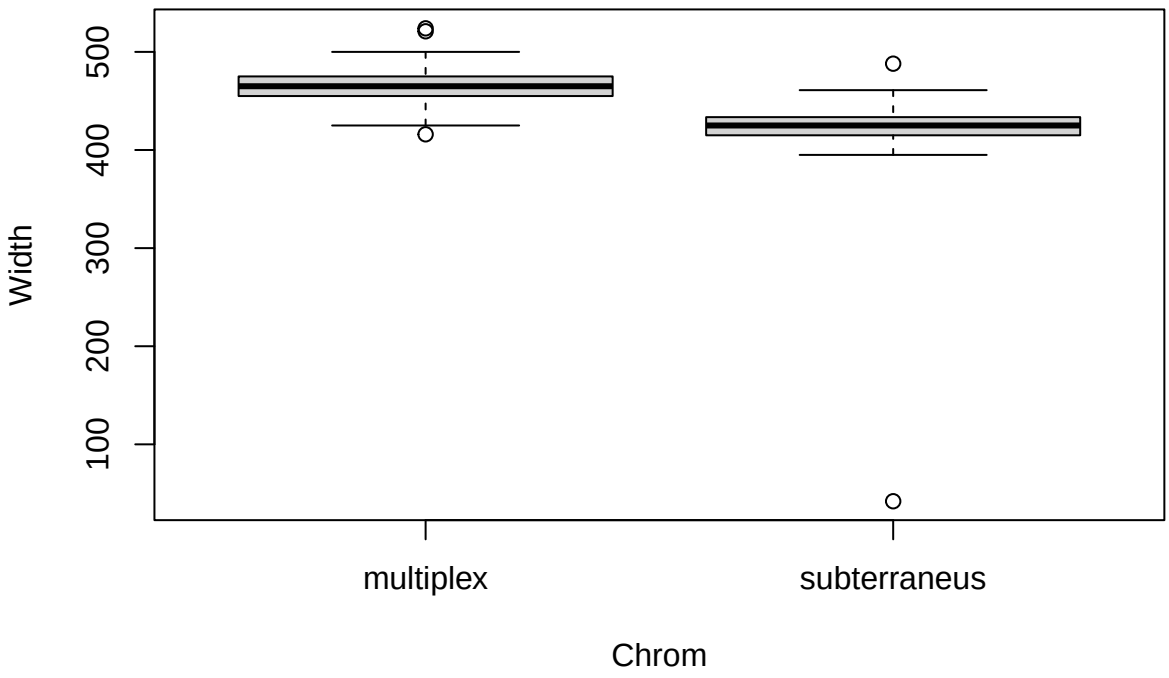
```
boxplot(Height ~ Chrom, data = data1)
```




```
data2 <- training_data[-c(13, 86, 33, 54),]
boxplot(Height ~ Chrom, data = data2)
```

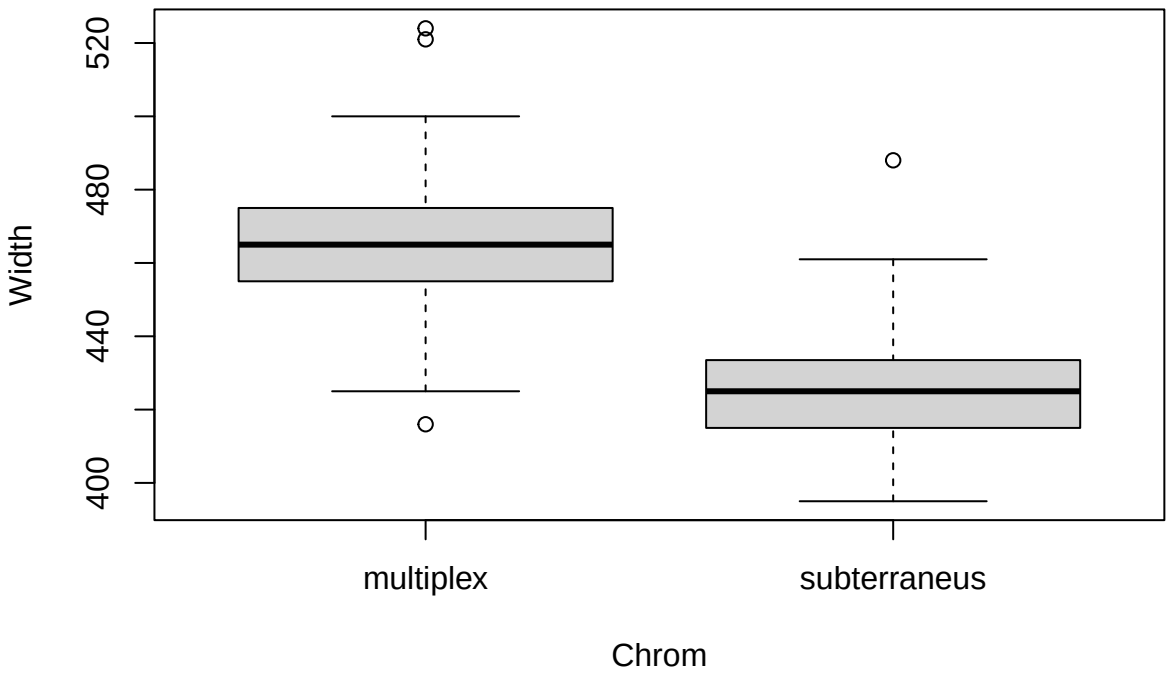



```
boxplot(Width ~ Chrom, data = data2)
```



An outlier in row 2 remained, so in all five rows of data were removed.

```
data3 <- training_data[-c(13, 86, 33, 54, 2),]
boxplot(Width ~ Chrom, data = data3)
```



I added a column to delineate multiplex from subterraneus.

```
data3$Chrom2 <- ifelse(data3$Chrom == "multiplex", 1, 0)
```

And performed a regression to look at GLM information and evaluate.

3. Explain your GLM and assess the quality of the fit with the classified observations. Use Cross Validation to predict the accuracy of your model.

I ran the GLM package to evaluate the clean data and note any significant correlations. Height had a significant impact on categorization, and width had an impact, but less of one.

```
regress1 <- glm(Chrom2 ~ Con + Height + Width, data = data3, family = "binomial")
summary(regress1)
```

```
##
## Call:
## glm(formula = Chrom2 ~ Con + Height + Width, family = "binomial",
##     data = data3)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48649  -0.38209  -0.05113   0.37603   2.58100
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -71.407839  16.252001  -4.394 1.11e-05 ***
## Con           0.001774   0.007593   0.234  0.81529
## Height        0.045805   0.016613   2.757  0.00583 **
## Width         0.070906   0.030968   2.290  0.02204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.823  on 84  degrees of freedom
## Residual deviance:  49.965  on 81  degrees of freedom
## AIC: 57.965
##
## Number of Fisher Scoring iterations: 6
```

This is where you actually test the bulk of the data to your training model. And create confusion matrix to show what percent you predicted correctly.

```
samp <- sample(c("train", "test"), nrow(data3), replace = T, prob = c(.8, .2))
train <- data3[samp == "train",]
test <- data3[samp == "test",]
model <- glm(Chrom2 ~ Con + Height + Width, data = train, family = "binomial")

test$pred_probs <- predict.glm(model,newdata=test,type="response")
pt <- table(test$Chrom2,test$pred_probs > 0.50)
## Specificity
pt[1,1]/sum(pt[1,])
```

```
## [1] 0.8666667
```

```
# Sensitivity
pt[2,2]/sum(pt[2,])
```

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

4. Provide a one-page write-up (excluding graphs, tables and figures) explaining your analysis of the dataset and your recommendations on the usefulness of your predictions. (Included here.)

5. Provide predictions for the unclassified observations.

   I added a predictions column based on the training data and applied it to the test data. Results below.

```
test_data$preds <- ifelse(predict.glm(regress1,newdata=test_data,type="response") > 0.5, "multiplex", "
head(test_data)
```

```
## # A tibble: 6 x 6
##   Index Chrom     Con Height Width preds
##   <dbl> <chr>   <dbl>  <dbl> <dbl> <chr>
## 1     1 unknown  2232    821   430 multiplex
## 2     2 unknown    NA    755   405 <NA>
## 3     3 unknown  2295     NA    NA <NA>
## 4     4 unknown  2355    842   490 multiplex
## 5     5 unknown  2335    814   481 multiplex
## 6     6 unknown  2355     NA   460 <NA>
```

In the test data, using my analysis and application of training model to the test data, there ended up being 72 samples labeled multiplex and 45 subterraneus.

```
install.packages("dplyr")
```

```
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

```
table(test_data["preds"])
```

```
## preds
##    multiplex subterraneus
##           72           45
```

6. As a secondary component provide annotated code that replicates your analysis. (Included here.)

Thank you!

-Anna Leisa Sauser