Question 1

Amin Baabol

11/21/2020

#Introduction “Microtus Subterraneus” and “Microtus 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 is 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 The dataset 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.

#Logistic Regression 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 /.The logit link function will then take the logarithm of the odds ratio and will increase with unrestricted range as the increases from 0 to 1.

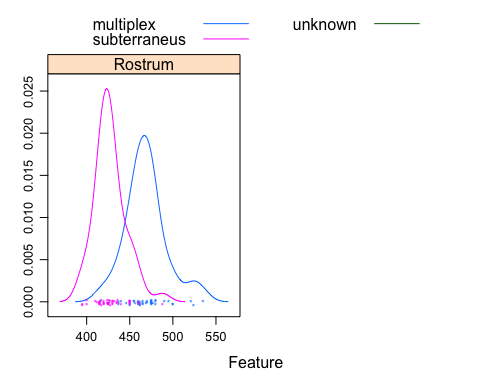
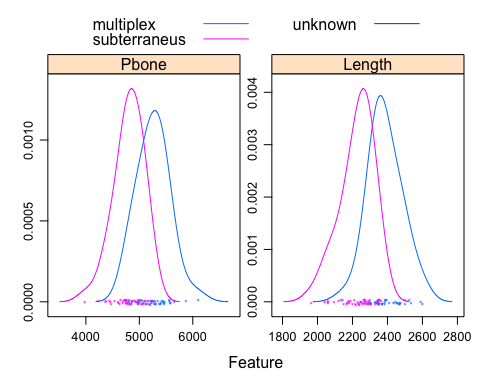
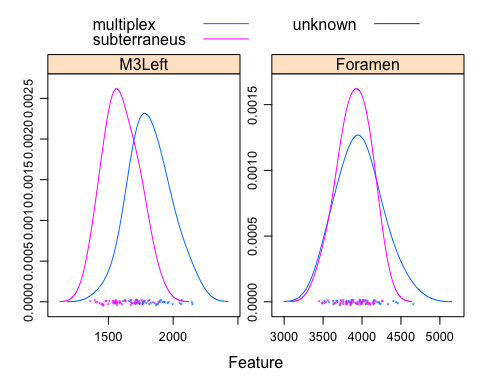
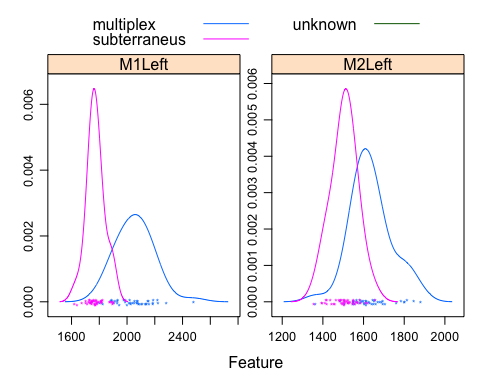
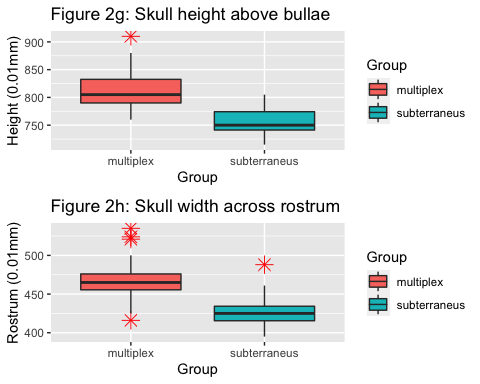
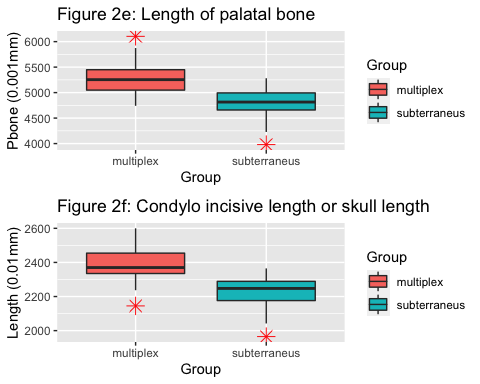
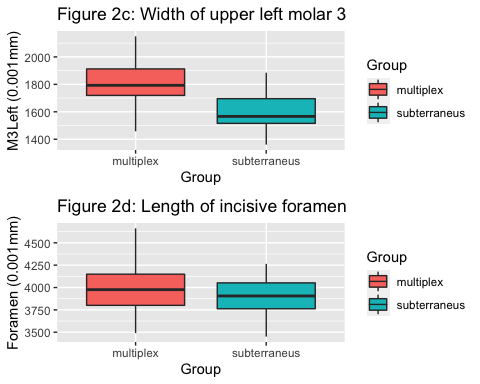
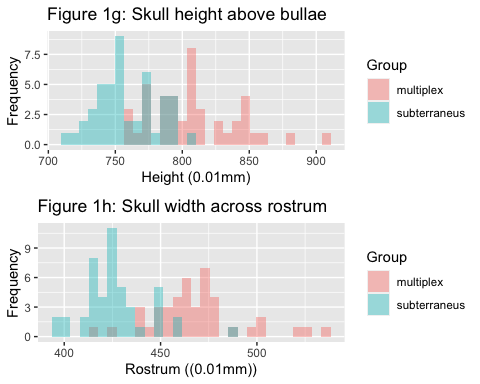
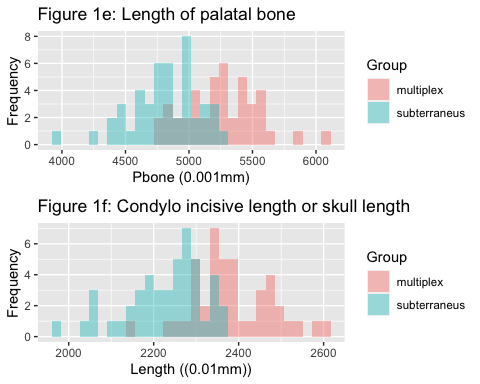
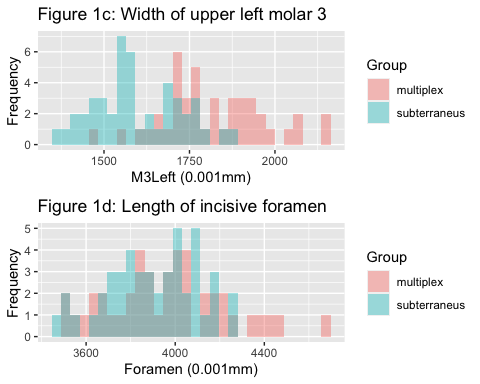
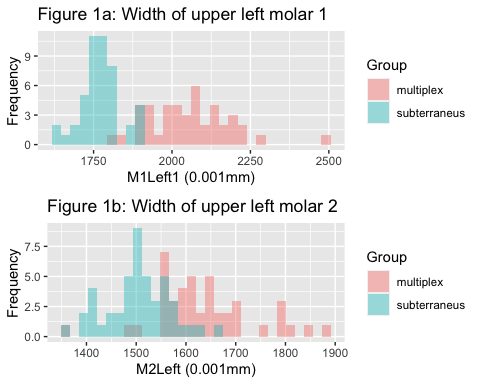
## Assumptions

The following assumptions are made in the process of a building binomial logistic regression model: 1.Predicted outcome is binary or discrete 2.The sample 2.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 Exploration We split our dataset into two subsets, the “known subset” will be used to train our models,which will then predict or classify the remaining “unknown subset”. Highly correlated variables will be removed from selection in order to decrease model complexity and to avoid multicollinearity and increasing standard error of the coefficient estimates. To perform multicollinearity diagnostics, we will use *ggpairs* from *GGally* package to create a correlation matrix.eliminated.

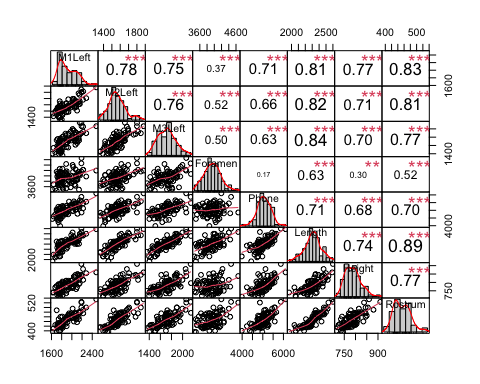
## Data Visualization

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.



## Correlation Matrix

In order to avoid over-fitting the logistic regression it is considered best best practice to eliminate highly correlated explanatory variables. Doing so will also assist us in preventing the possible presence of multicollinearity in the model. The correlation plot suggests that almost all of our predictors have highly significant correlations.To remove the highly correlated variables an arbitrary cut-off value of 75% was selected.This significantly reduced the predictor variables available for selection down to three(Figure 3): 1.M1Left - Width of upper left molar 1 (0.001mm) 2.Foramen - Length of incisive foramen (0.001mm) 3.Pbone - Length of palatal bone (0.001mm)



## 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.

## Group M1Left M2Left M3Left Foramen Pbone Length Height Rostrum  
## 1 1 2078 1649 1708 3868 5463 2355 805 475  
## 2 1 1929 1551 1550 3825 4741 2305 760 450  
## 3 1 1888 1613 1674 4440 4807 2388 775 460  
## 4 1 2020 1670 1829 3800 4974 2370 766 460  
## 5 1 2223 1814 1933 4222 5460 2470 815 475  
## 6 1 2190 1800 2066 4662 4860 2535 838 521  
## 7 1 2136 1640 1767 4070 5372 2385 815 480  
## 8 1 2150 1761 1859 4053 5231 2445 840 480  
## 9 1 2040 1694 1958 3977 5579 2435 835 440  
## 10 1 2052 1551 1712 3877 5401 2330 830 475  
## 11 1 2282 1706 1896 3976 5560 2500 855 500  
## 12 1 1892 1626 1763 3538 5149 2270 810 446  
## 13 1 1977 1556 1935 3576 5346 2330 785 462  
## 14 1 2220 1680 2054 4226 5130 2465 880 490  
## 15 1 2070 1604 1616 3633 5037 2345 845 475  
## 16 1 2000 1602 1818 3997 5304 2410 790 460  
## 17 1 2140 1612 1719 3490 5254 2305 790 450  
## 18 1 2084 1565 1793 3834 5078 2345 760 450  
## 19 1 2072 1651 1772 3970 5402 2396 804 462  
## 20 1 2132 1784 1875 4150 5422 2390 845 460  
## 21 1 1826 1548 1815 3519 5230 2250 800 425  
## 22 1 2073 1588 1919 4239 5203 2385 790 475  
## 23 1 2187 1801 2145 4464 5874 2600 910 524  
## 24 1 1802 1363 1458 3631 4842 2145 760 416  
## 25 1 2054 1569 1745 3678 5445 2305 791 462  
## 26 1 2479 1880 2065 4195 6104 2590 860 535  
## 27 1 2102 1506 1660 3871 5212 2300 772 437  
## 28 1 2158 1612 1869 4015 5652 2500 828 480  
## 29 1 1907 1549 1672 4050 5307 2350 770 456  
## 30 1 2084 1660 1906 4000 5061 2355 805 465  
## 31 1 1987 1592 1720 3741 5245 2475 810 470  
## 32 1 1933 1486 1742 4007 5032 2345 810 465  
## 33 1 1914 1583 1722 3677 4871 2237 805 437  
## 34 1 2015 1695 1997 4404 5453 2525 815 495  
## 35 1 1930 1688 1883 3941 5004 2370 795 469  
## 36 1 2155 1656 2150 4070 5473 2457 796 477  
## 37 1 1988 1599 1779 3856 5165 2352 770 475  
## 38 1 2027 1645 1966 4334 5293 2452 775 470  
## 39 1 2023 1612 1781 4148 4940 2340 796 455  
## 40 1 1885 1549 1628 3718 5286 2300 810 455  
## 41 1 1945 1580 1739 3801 5567 2370 800 475  
## 42 1 2186 1847 1896 4160 5587 2470 845 500  
## 43 1 2110 1631 1703 3856 4773 2350 850 465  
## 44 0 1888 1548 1763 4112 4814 2350 735 450  
## 45 0 1898 1568 1734 4169 4919 2285 750 420  
## 46 0 1735 1534 1566 3947 4773 2170 738 415  
## 47 0 1746 1394 1397 3657 4771 2060 720 415  
## 48 0 1734 1495 1561 3859 5229 2275 785 417  
## 49 0 1741 1530 1683 3999 4745 2330 790 450  
## 50 0 1746 1562 1456 3807 5108 2260 760 426  
## 51 0 1722 1558 1757 4097 4379 2290 750 432  
## 52 0 1873 1524 1885 3921 5007 2340 795 450  
## 53 0 1738 1419 1634 4039 4228 2270 771 420  
## 54 0 1731 1546 1560 3764 4866 2175 755 424  
## 55 0 1815 1436 1361 3728 4911 2150 750 412  
## 56 0 1790 1524 1606 3890 4700 2189 770 427  
## 57 0 1814 1454 1672 3890 5282 2275 795 425  
## 58 0 1819 1506 1809 3564 5062 2290 790 435  
## 59 0 1814 1550 1552 4265 4801 2298 776 440  
## 60 0 1773 1355 1447 3717 4649 2135 745 415  
## 61 0 1783 1465 1487 4141 4459 2240 760 415  
## 62 0 1762 1657 1717 4262 4982 2270 750 435  
## 63 0 1766 1585 1557 3805 4474 2235 731 430  
## 64 0 1823 1504 1591 3928 4611 2275 725 425  
## 65 0 1795 1487 1478 3762 5000 2200 775 415  
## 66 0 1702 1522 1725 4155 5065 2350 779 488  
## 67 0 1755 1517 1536 4098 4634 2265 746 420  
## 68 0 1811 1611 1537 4081 4998 2365 740 461  
## 69 0 1776 1582 1715 3989 5118 2342 788 460  
## 70 0 1674 1491 1433 3521 4724 2042 770 425  
## 71 0 1770 1490 1586 3762 4971 2250 740 425  
## 72 0 1902 1499 1680 4056 5178 2300 755 450  
## 73 0 1814 1510 1677 3856 4689 2245 805 430  
## 74 0 1728 1505 1544 3726 4746 2120 750 420  
## 75 0 1714 1525 1590 3973 4957 2230 725 425  
## 76 0 1895 1480 1561 3991 4816 2210 772 450  
## 77 0 1758 1507 1631 3852 4979 2221 765 430  
## 78 0 1640 1416 1542 3687 4601 2095 740 410  
## 79 0 1770 1621 1567 4156 4773 2286 745 436  
## 80 0 1746 1419 1700 4021 4368 2182 735 400  
## 81 0 1784 1502 1417 3959 4815 2168 750 424  
## 82 0 1781 1504 1731 3649 5104 2260 748 427  
## 83 0 1770 1396 1509 3864 3980 2061 715 400  
## 84 0 1702 1443 1500 3451 4977 2060 745 395  
## 85 0 1779 1572 1771 4016 5199 2355 792 425  
## 86 0 1747 1411 1566 3803 4537 2180 750 417  
## 87 0 1878 1549 1844 4078 4747 2295 795 430  
## 88 0 1619 1458 1402 3492 4439 1965 740 395  
## 89 0 1749 1482 1462 3797 4855 2218 765 415

## Analysis of Deviance Table  
##   
## Model 1: Group ~ 1  
## Model 2: Group ~ M1Left  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 88 123.279   
## 2 87 28.517 1 94.762 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Call:  
## glm(formula = Group ~ M1Left + Foramen, family = binomial(link = "logit"),   
## data = known.subset)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.28036 -0.09923 -0.01058 0.01788 2.49687   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -62.804452 20.661080 -3.040 0.00237 \*\*   
## M1Left 0.047246 0.014091 3.353 0.00080 \*\*\*  
## Foramen -0.006637 0.003192 -2.079 0.03758 \*   
## ---  
## 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: 22.049 on 86 degrees of freedom  
## AIC: 28.049  
##   
## Number of Fisher Scoring iterations: 8

## Analysis of Deviance Table  
##   
## Model 1: Group ~ M1Left  
## Model 2: Group ~ M1Left + Foramen  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 87 28.517   
## 2 86 22.049 1 6.468 0.01098 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table  
##   
## Model 1: Group ~ M1Left + Foramen  
## Model 2: Group ~ M1Left + Foramen + Pbone  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 86 22.049   
## 2 85 21.738 1 0.31124 0.5769

## Analysis of Deviance Table  
##   
## Model 1: Group ~ M1Left + Foramen  
## Model 2: Group ~ M1Left + Foramen + Pbone + M3Left  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 86 22.049   
## 2 84 21.070 2 0.97946 0.6128

## Analysis of Deviance Table  
##   
## Model 1: Group ~ M1Left + Foramen  
## Model 2: Group ~ M1Left + Foramen + Pbone + M3Left + Height  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 86 22.049   
## 2 83 18.493 3 3.5561 0.3136

## Analysis of Deviance Table  
##   
## Model 1: Group ~ M1Left + Foramen  
## Model 2: Group ~ M1Left + M2Left + M3Left + Foramen + Pbone + Length +   
## Height + Rostrum  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 86 22.049   
## 2 80 14.962 6 7.0871 0.3129

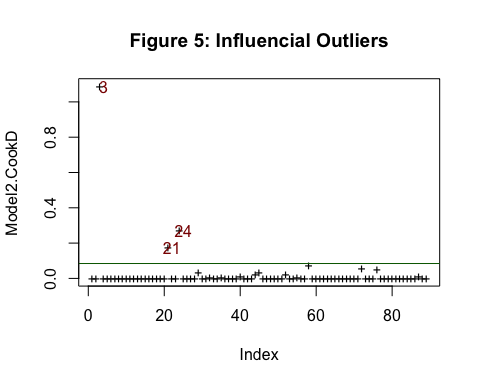
## Analysis of Deviance Table  
##   
## Model 1: Group ~ 1  
## Model 2: Group ~ M1Left  
## Model 3: Group ~ M1Left + Foramen  
## Model 4: Group ~ M1Left + Foramen + Pbone  
## Model 5: Group ~ M1Left + Foramen + Pbone + M3Left  
## Model 6: Group ~ M1Left + Foramen + Pbone + M3Left + Height  
## Model 7: Group ~ M1Left + M2Left + M3Left + Foramen + Pbone + Length +   
## Height + Rostrum  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 88 123.279   
## 2 87 28.517 1 94.762 < 2e-16 \*\*\*  
## 3 86 22.049 1 6.468 0.01098 \*   
## 4 85 21.738 1 0.311 0.57692   
## 5 84 21.070 1 0.668 0.41367   
## 6 83 18.493 1 2.577 0.10845   
## 7 80 14.962 3 3.531 0.31677   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 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

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



## [1] 3.679744

## Model Performance

### Cross Validation

Having a small training data set creates performance uncertainty in our model, therefore,it is very crucial to perform at least one method of re-sampling to properly validate our models. Model statistics Summary Table: ||| || ——–|————–|——–|——|—–|————— | 3.68% | 4.71% |28.04 |35.51|2 predictors

## [1] 4.70638

## [1] 28.04904

## [1] 35.51495

##Prediction Predicting the remaining unclassified 199 specimens.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000033 0.0340634 0.9758355 0.6180224 0.9999450 1.0000000

## Classified\_Samples  
## multiplex subterraneous   
## 120 79

#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.