23. Binomial ANOVA

This is a bonus lab. You are not required to know this information for the final exam.

Binomial data occurs when your data has two mutually-exclusive classes (data cannot belong to both classes). Some common examples of binomial responses include presence/absence of species, survival/mortality of units (hopefully plants, not people!), simple yes/no responses, pass/fail, infected/uninfected with disease, etc.

We have already discussed tests suitable for binomial data, but for the cases where we have 2 or more predictor variables we can also run an ANOVA using the output from a generalized linear model referencing logistic regression and the binomial distribution.

23.1 Presence-Absence (i.e. binomial response) Data

First, we will look at some data and see if we can see some trends in the probabilities. Remember that we're going to reduce our binomial response to a probability of one response or the other.

- Download the “LogReg.csv” data from the class website, import it into R and attach the data file. The data consists of several environmental predictors, including mean annual temperature (MAT), mean annual precipitation (MAP), Elevation, and a categorical predictor of slope aspect (north or south). The response variable is a presence-absence (1 or 0) of lodgepole pine (Pine_PA).
- Have a look at the relationship between MAT and pine presence. Then look at MAP and pine presence. Which variable looks like it might be a better predictor of pine presence?
  ```r
  plot(Pine_PA~MAT)
  plot(Pine_PA~MAP)
  ```

23.2 Binomial ANOVA

- A logistic regression works just like a simple or multiple linear regression. It is capable of handling continuous or categorical independent variables (predictors). To run a binomial ANOVA you must build your logistic model using the `glm()` command, which is a bit more robust than the alternative linear model `lm()` function. To indicate that you have a binomial response, we must tell R the family of the model. Have a look at the AIC.
  ```r
  lr.model = glm(Pine_PA~MAT+MAP, family="binomial")
  summary(lr.model)
  AIC(lr.model)
  ```
- The next step is to use the model output the `anova()` function. However unlike the parametric option, here we must specify that we want to calculate p-values using a chi-squared test and chi-squared distribution rather than the F-distribution (`anova()` default) which is reserved for parametric statistics.
  ```r
  lr.anova= anova(lr.model, test="Chisq")
  lr.anova #ANOVA output
  ```
- The output from this ANOVA is similar to the parametric option, however rather than the sum-of-squares and means sum-of-squares displayed, now we have the resulting deviance from each of the parameters. Remember deviance is a measure of the lack of fit between the model and the data with larger values indicating poorer fit
- The p-values are calculated using the chi-squared distribution, but like the parametric alternative they indicate whether or not each of the predictors has a significant effect on the probability of achieving a "success" (value of 1).
- If you have a biological reason to do so, you can also determine if there are interactions in our data by including * in our model statement
```
  lr.model = glm(Pine_PA~ MAT*MAP,family="binomial")
  summary(lr.model2)
  AIC(lr.model2)

  lr.anova2=anova(lr.model2, test="Chisq")
  lr.anova2 #ANOVA output
```

- If you find there are significant interactions, this could indicate collinearity among your predictors or it could indicate the effect of the predictor on the response is more complex (just like when we look at interactions in a parametric ANOVA). You could then follow this up with pairwise z-tests for proportions with adjusted p-values for multiple comparisons to determine the effects.

- More importantly the ANOVA could give you a better indication of predictor variables to include in your logistic model. The interaction between predictor variables might be a better predictor for your model than the raw variables themselves (indicated by significant p-values). You could improve your overall model fit by including the interaction between 2 variables which will reduce complexity by reducing the number of parameters. Look at the AIC values from the models above, which one could be considered a better logistic model?