---

title: "report for lab 6"

author: "Harry O'Brien"

date: "3/12/2021"

The Data

The data has 11 variables total. One of which is the ID which was quickly discarded, along with all row entries with NA values. 1 of the variables, class_tumor, was the outcome variable, with the other 9 variables being correlated to this. To clean up the data all entries were converted to integers, and class_tumor was converted to a 0-1 binary.

Training subsets

The training and testing subsets were created by randomly splitting the data into 2 subsets, 80% of which was the training subset and the remainder being the test subset.

The Model

The model used the glm function as this was the subject of the last lecture. We can see from the p values on summary(model) that some variables were particularly correlated with the diagnoses, specifically clump_thick, bare_nuclei10, and bare_nuclei4. I trained the model on the train subset and tested on the test subset. Then I found the are under the ROC curve and, using this info, identified 4 to be the optimal threshold for minimizing false positives/negatives.