# STAT-S 432: Homework 4

*Due March 29, 2019 by 11:59 PM*

**Instructions:** You must submit this homework by pushing a file named "hw4.Rmd" file to your team's repo. Note that that is the **only** file you will be allowed to push. Commit early and often.

**Data named n90_pol.csv**

The data set `n90_pol.csv` contains information on 90 university students who participated in a psychological experiment designed to look for relationships between the size of different regions of the brain and political views. The variables `amygdala`and `acc` indicate the volume of two particular brain regions known to be involved in emotions and decision-making, the amygdala and the anterior cingulate cortex (ACC); more exactly, these are residuals from the predicted volume, after adjusting for height, sex, and similar anatomical variables. The variable orientation gives the subjects' scores on a five-point scale from 1 (very conservative) to 5 (very liberal). `orientation` is an ordinal but not a metric variable, so scores of 1 and 2 are not necessarily as far apart as scores of 2 and 3.

1. Creating a binary response variable. (5 points)
    - Create a vector, `conservative`, which is 1 when the subject's `orientation` is less than or equal to 2, and 0 otherwise.
    - Explain why the cut-off was put at an `orientation` score of 2 as opposed to some other cut-off.
    - Use code to check that your `conservative` vector has the proper values without manually examining all 90 entries.
    - Add the `conservative` vector that you created to the existing data frame.

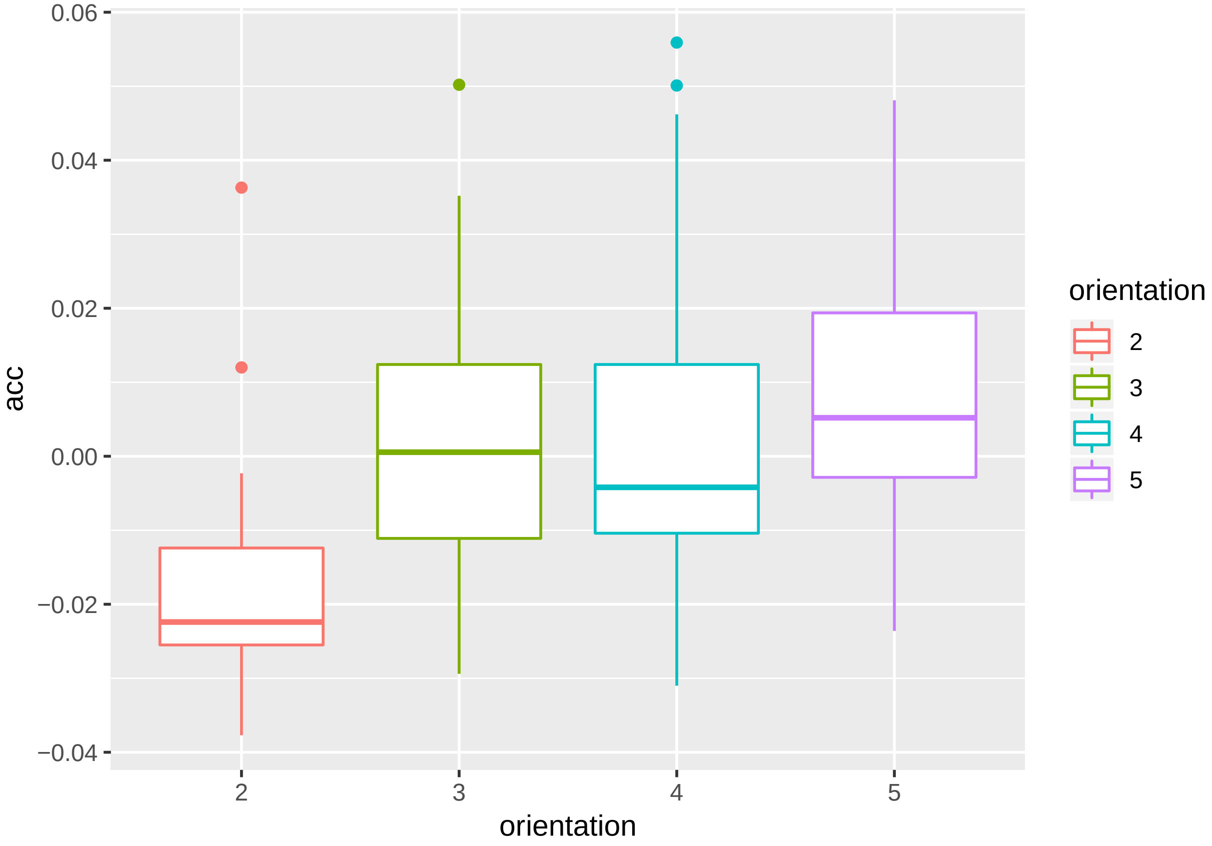
A binary variable was created to divide the indiviudals into two groups, conservative and not. The orientation of the students was used to make this categorization. Specifically, scores of 2 or less were used to classifiy someone as a conservative. The scoring range is from 1 to 5 with 1 being most conservative and 5 being most liberal. This implies a 3 would be neutral. Therefore someone would be classified if their score was a 1 or 2. Scores of 3, 4, or 5 would indicate that someone is at least neutral (not conservative) or liberal (definitely not conservative).

2. Perform exploratory data analysis on the data. (7 points)
    - Examine the univariate distributions of the varaibles. Describe the distributions.
    - Examine bivaraite relations. This should include the following:
        - how the ACC and amygdala volumes change across the different orientation scores.
        - how the ACC and amygdala volumes relate to each other.
        - how the distribution of both ACC and amygdala brain volumes change whether someone is classified as a "conservative" or not.
    - Create a scatter plot of `amygdala` versus `acc`. Use shape and color to show whether an individual is a "conservative" or not on the scatterplot. Explain how the amygdala and ACC volumes may be related to the chance that a randomly selected student is a "conservative" or not.

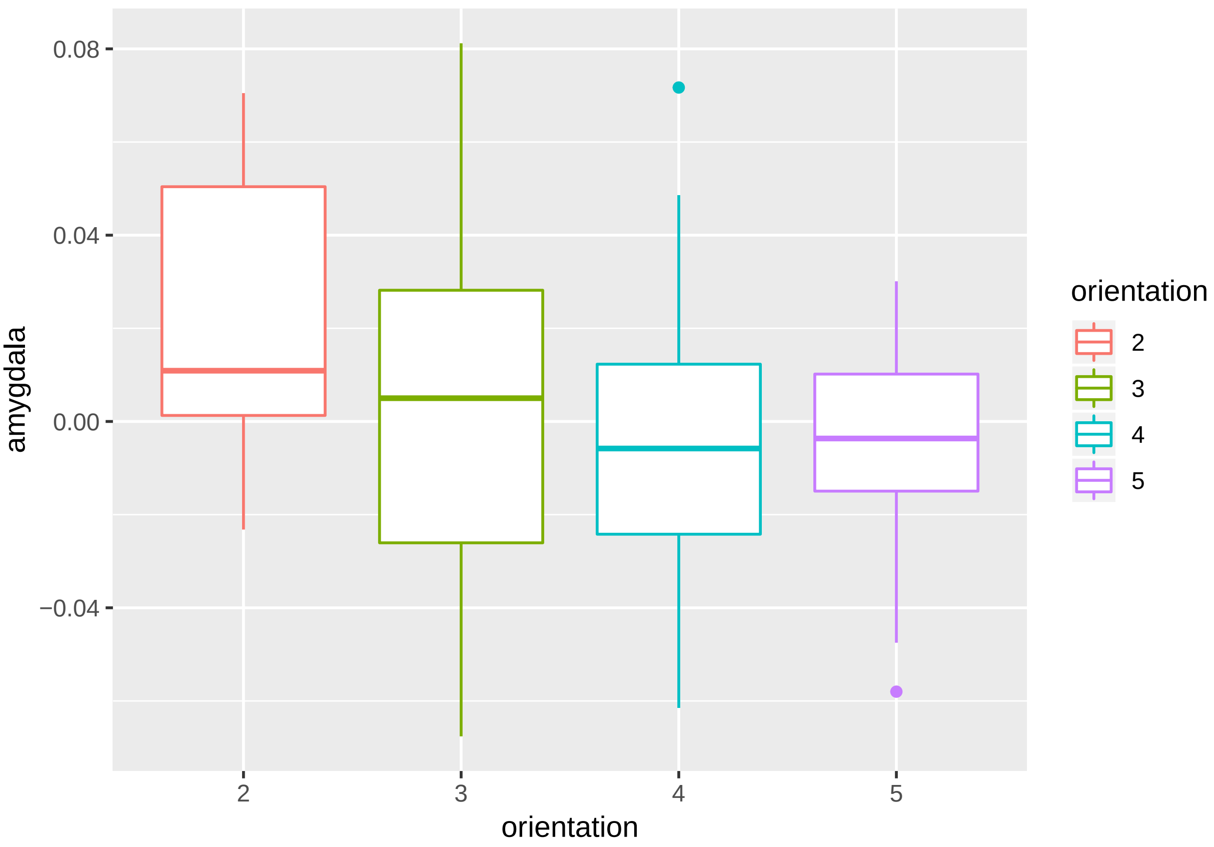
There were a total of 90 observations in the data. Each observation is information from a single student. Variables recorded from the students are their political orientation on a 1 to 5 scale, their ACC brain volume (residual), and their amygdala brain volume (residual). Additionally a variable was coded to separate the students into conservatives and non-conservatives based on their orientation being 2 or below for conservative and 3 or above for non-conservatives. It should be noted that no individuals had an orientation below 2. This may be a product of students typically having a more liberal skew than older age groups.

The objective will be to see how well the brain volumes will correlate with someone's conservative status. Starting with, we will see how the brain volumes relate to the orientation score. Then we will move on to examining the relationship of brain volumes and conservative status.

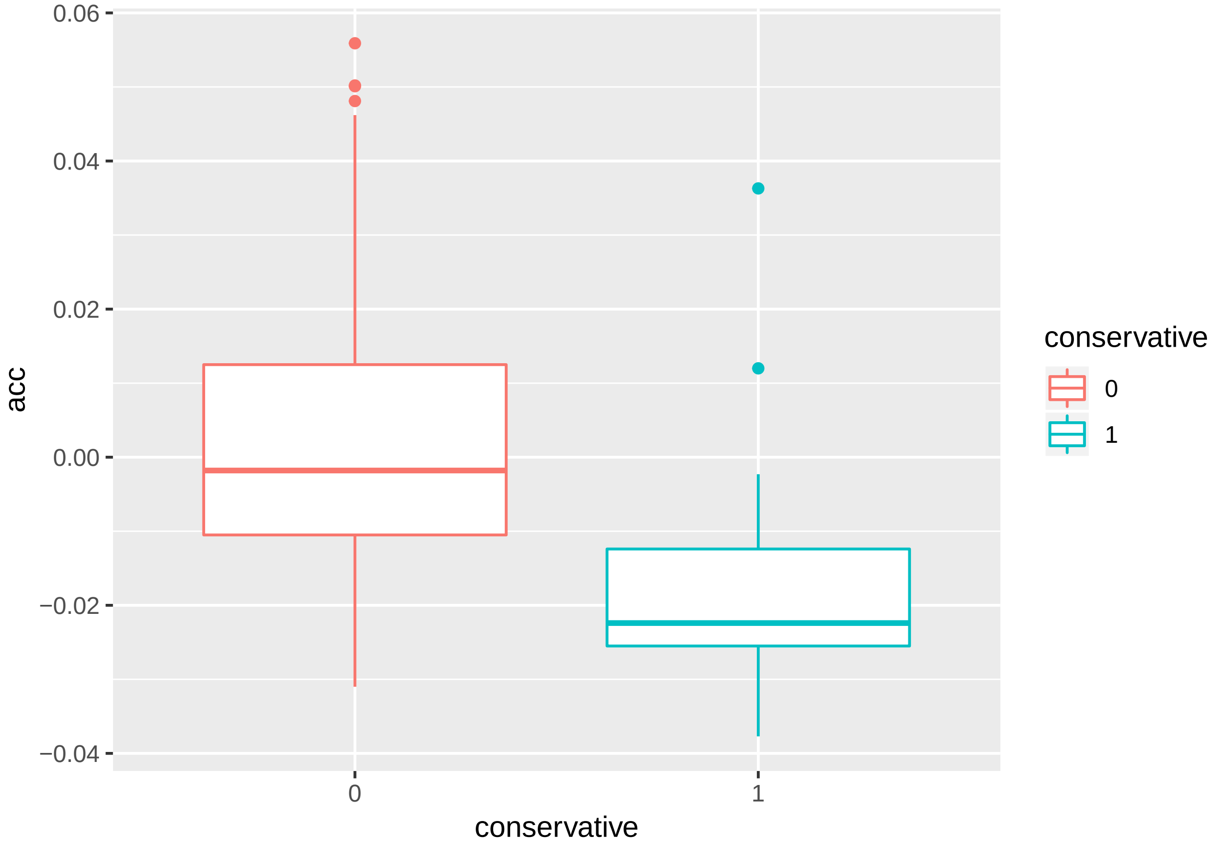
The first figure below shows boxplots that display the ACC brain volumes broken up the orientation scores. The most striking feature is that the distribution of brain volumes for an orientation score of 2 are noticeably lower than the ACC brain volumes for any other orientation score. Also, the ACC brain volumes show noticeably less variability in the group with an orientation score of 2 than the other groups except for two depicted outliers. With the inclusion of the outliers, the distribution of ACC volumes of those with an orientation of 2 is also much more right skewed than the distribution of the other The distribution of ACC brain volumes among those with orientation scores of 3 or above have fairly similar distributions in terms of the quartiles and spread. An orientation score of 5 is associated with slightly higher brain volumes compared to the other groups, but this is not a very noticeable difference among the three groups with scores of 3,4 and 5.



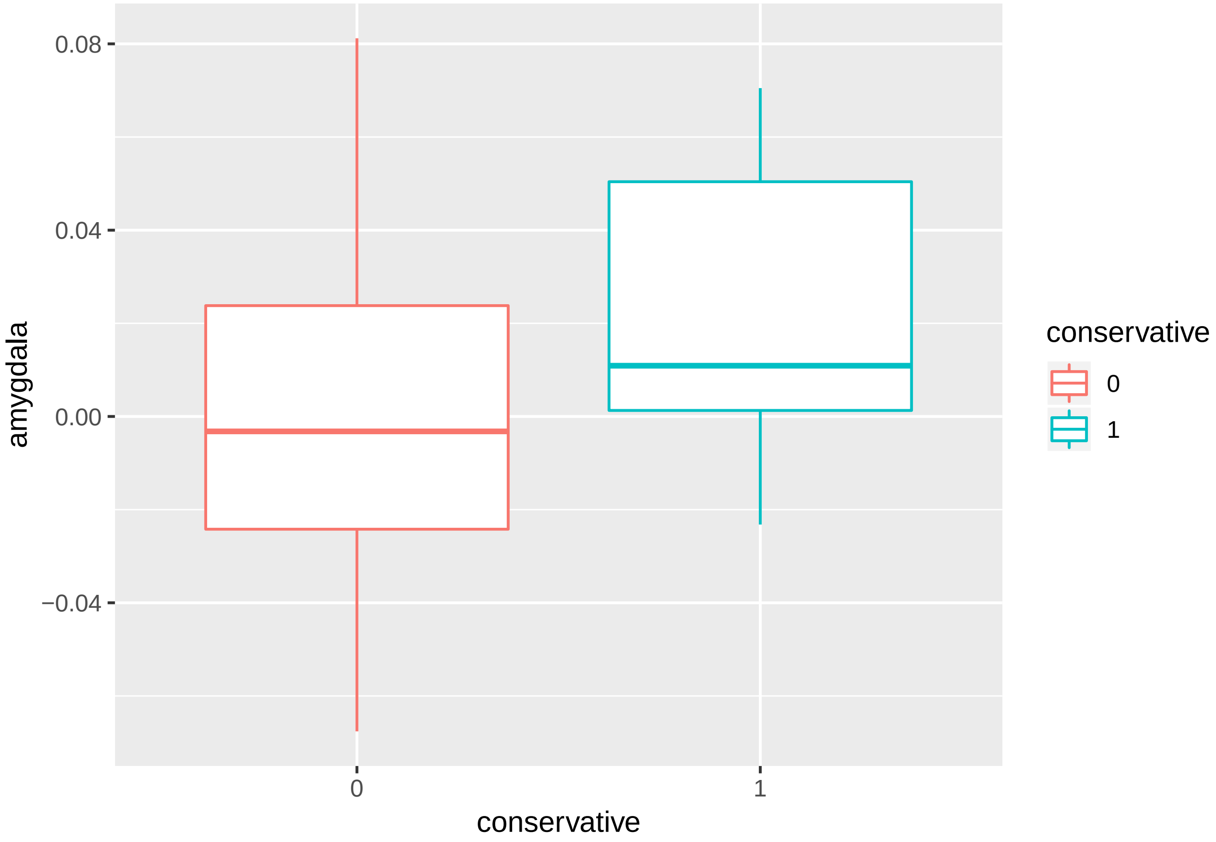
In the next figure, we see the amygdala brain volumes across the different orientation scores. The medians for all four groups are relatively similar, the main differences are in shape and spread. The brain volumes in group 2 is more right skewed than the other groups but does not contain the highest scores overall. The distribution in the other groups are more symmetric with decreasing spread as the orientation score increases. Overall it appears that the ACC brain volume is a much better variable for separating the different groups because the ACC brain volumes are shifted below the brain volumes of the other groups.



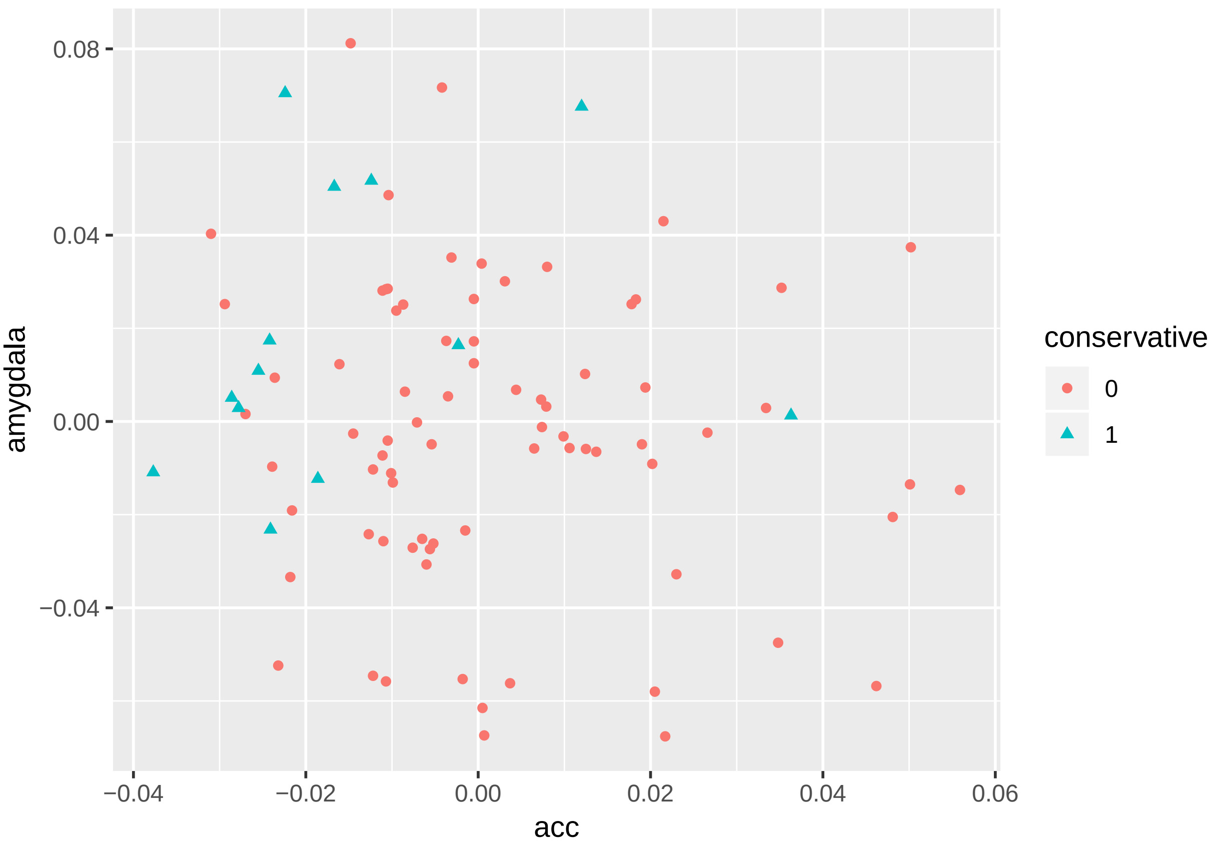
In the next figure below, boxplots showing the distribution of ACC volumes separated by conservative status are displayed. Conservatives have lower ACC volumes than non-conservatives as well as lower variability. The ACC volume appears to be a decent variable for identifying conservative status. Though there is quite a bit of overlap of the range of brain volumes between the two groups.



The proceeding plot shows amygdala volumes for conservatives versus non-conservatives. The distribution for conservatives is more right skewed and non-conservatives have volumes that extend much lower than conservatives. Given that the central point of both distributions is very similar and the groups have overlapping ranges, the amygdala brain volume will not be as good of a classifying variable than the ACC volume.



Finally, a scatter plot of the two potential predictors is presented below to assess the correlation between the ACC volumes (x-axis) and amydala volumes (y-axis) Additionally, colors and shapes distinguish between conservatives and non-conservatives. The scattering of points ignorive conservative status is completely random indicating there is not much if any relationship between the variables in the joint distribution. There is some separation between conservatives and non-conservatives with ACC volume below 0, though there are still fairly many non-nonconservatives in this range as well. This overlap means that it will difficult to discriminate between the two groups. As observed earlier, there doesn't seem to be a point in amygdala volumes that separates conservatives versus non-conservatives clearly.



3. Fit a logisitic regression model of `conservative` on `amygdala` and `acc`. (8 points)
   - Provide a kable-table of the coefficients, their standard errors, test-statistics, and p-values.
   - Summarize the table including a an interpretation of how the predictor varaibles affect the *odd ratio* (not the log-odds ratio).
   - Based on your table, is the model an effective way of classifying a randomly chosen student as a "conservative" or not. (Ignore the fact that we would have to get their brain volumes…)

A logistic regression model was used to attempt to predict the conservative status of individuals. Both ACC and amygdala volumes were the predictors in the model. A summary table is presented below. For interpretation purposes, the brain volume residuals were multiplied by 100 so that when interpreations of model coefficients are given, we can give the interpretations in terms increases of respective brain volumes by 1/100ths.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -2.458 | 0.491 | -5.01 | 0.000 |
| I(acc * 100) | -0.655 | 0.252 | -2.60 | 0.009 |
| I(amygdala * 100) | 0.220 | 0.109 | 2.02 | 0.044 |

The resulting formula takes the form

$$\log\left(\frac{p(\vec{x})}{1-p(\vec{x})}\right) = -2.458 - 0.655acc + 0.22amygdala$$

This indicates that increases in ACC volume are associated with decreases in the probability that someone is a conservative, and increases in amygdala volume is associated with an increase in the probability. Specifically, an increase of ACC volume by 1/100th of a unit would correspond to a -48.056% change in the odds that someone is a conservative. And an increae in the amygdala volume by 1/100th of unit is predicted to yield a 24.608 change in the odds that someone is a conservative.

Though, the hypothesis tests should be used with extreme caution, we have significant results for ACC brain volume and somewhat weak results for the significance of the amygdala volume. These results correspond with examination of the variables in the EDA earlier.

It would appear that the model has at least some level of effectiveness for predicting the odds of someone being conservative. This effective relies mostly on the ACC brain volume. Though the results are based a fairly small sample size so it remains to be seen if the model is actually effective. This will be examined through misclassification rates.

4. We want to use the logistic regression model to classify individuals as conservative or not. (5 points)
    - Find predictions for each subject. Use the probability 0.5 as a cut-off for your classification.
    - What fraction of subjects are mis-classified?
    - What fraction of subjects would be mis-classified by “predicting” that none of them are conservative?
    - Based on this information, is this an effective model or not?

To examine the effectiveness of the model, predictions were obtained for the observations in the dataset. To do this, the fitted probability of being a conservative was obtained for each observation in the dataset, and any individual with a fitted probability $\geq 0.5$ was classified as a conservative.

The results were that 15 observations were misclassified, which out of the 90 observations leads to a misclassiication rate of 16.667%. This is not a terrible misclassification rate, but since the objective is to predict if an individual is a conservative, we should be concerned with the sensitivity of the model. The sensitivity would be the proportion of conservatives detected by the model out of the the total number of observed conservatives. This is reported in the confusion matrix below.

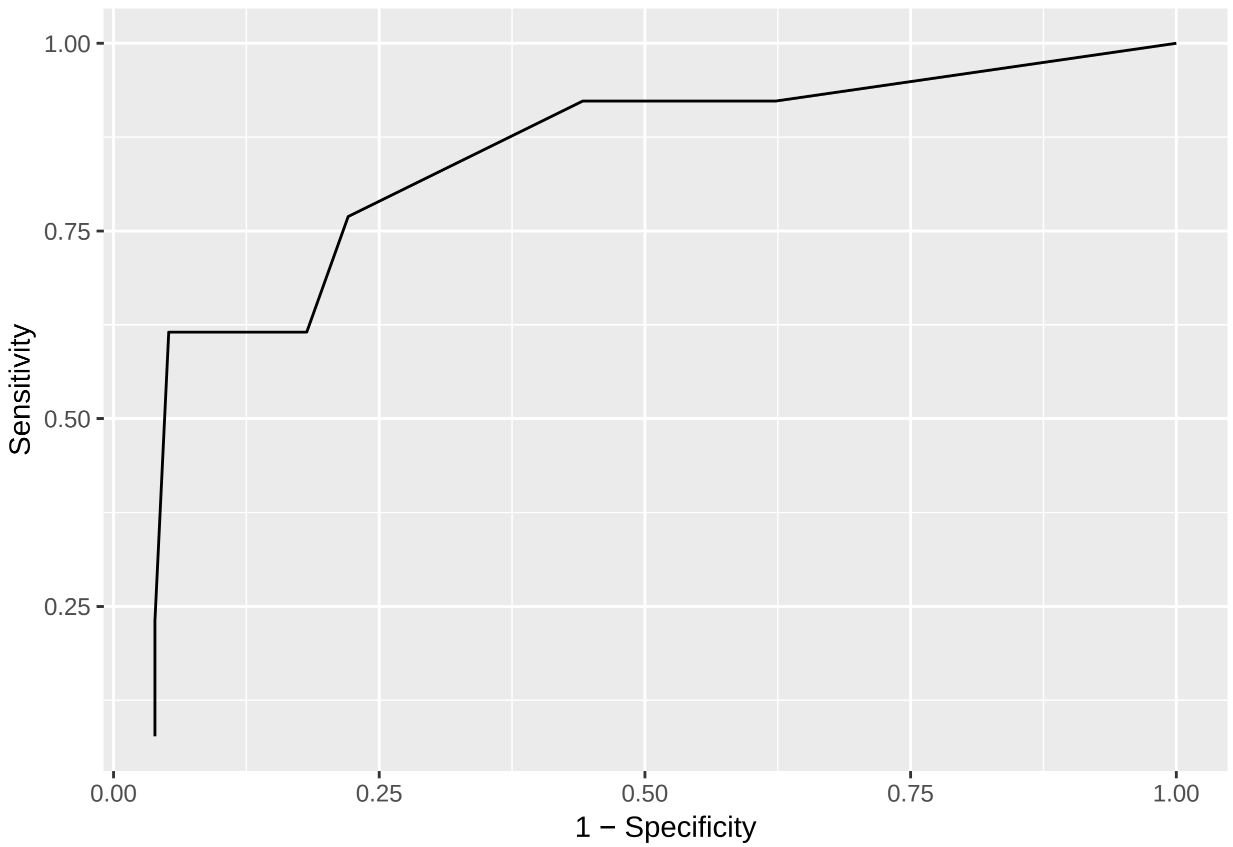
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 74 12
##          1  3  1
##
##                Accuracy : 0.833
##                  95% CI : (0.74, 0.904)
##     No Information Rate : 0.856
##     P-Value [Acc > NIR] : 0.7780
##
##                   Kappa : 0.053
##  Mcnemar's Test P-Value : 0.0389
##
##             Sensitivity : 0.0769
##             Specificity : 0.9610
##          Pos Pred Value : 0.2500
##          Neg Pred Value : 0.8605
##              Prevalence : 0.1444
##          Detection Rate : 0.0111
##    Detection Prevalence : 0.0444
##       Balanced Accuracy : 0.5190
##
##        'Positive' Class : 1
##
```

Out of the 13 conservatives in the data, we detected only 1 of them using the model! The sensitivity is abysmally low at 7.69%. I suppose we can take consolation in the specificty (proportion of true negatives in the model) is 96.1%. As a demonstration of how poorly the model works, we can compare our results to a model that predicts that no individuals are conservative. This leads to a misclassifcation rate or 14.444%. Which is actually lower than the logistic regression model! The logistic regression model is completely ineffective.

## Extra bit. Lower cutoff

If we are concerned mainly with the sensitivity of the model, it may be beneficial to decrease the cutoff for classifying someone as a conervative. The table below shows the sensitivity, specificity, and overall misclassification rate for a sequence of cutoffs. Choice of cutoff is relative to situation depending on how much a 'cost' one puts on misclassifying conservatives.

| Cutoff Probability | Sensitivity | Specificity | Misclassification Rate |
|---:|---:|---:|---:|
| 0.50 | 0.077 | 0.961 | 0.167 |
| 0.45 | 0.077 | 0.961 | 0.167 |
| 0.40 | 0.231 | 0.961 | 0.144 |
| 0.35 | 0.615 | 0.948 | 0.100 |
| 0.30 | 0.615 | 0.909 | 0.133 |
| 0.25 | 0.615 | 0.909 | 0.133 |
| 0.20 | 0.615 | 0.818 | 0.211 |
| 0.15 | 0.769 | 0.779 | 0.222 |
| 0.10 | 0.923 | 0.558 | 0.389 |
| 0.05 | 0.923 | 0.377 | 0.544 |
| 0.00 | 1.000 | 0.000 | 0.856 |



### Rubric

Scoring will be done according to the following criteria.

Each question or bullet point has been addressed. Questions are answered accurately. Any plot or table produced is explained and relationships are described accurately.

If document does not knit (assuming required packages are installed) -5

Problem 1 (5 points)

Problem 2 (7 points)

Problem 3 (8 points)

Problem 4 (5 points)