Predicting Heart Disease

Casey Freimund, Alyssa Phippen, Samuel Vanderzee, Kelli Walter

10/30/2019

Introduction

Given the data set including information on 14 different variables for 303 patients, we wanted to determine which of the variables were important in predicting heart disease in patients. In order to make these predictions, we started by exploring the data to see where important relationships lie and to see what groups the data formed in order to create a model. We created a model to predict the chances of heart disease based on the variables we deemed significant from earlier data exploration and model exploration.

Exploring the Data

During our data exploration, we found missing values in the "ca" column (number of major vessels covered by fluoroscopy) as well as in the "thal" column. Instead of deleting these values, we decided to replace these entries with the median value of the values in each of these columns.

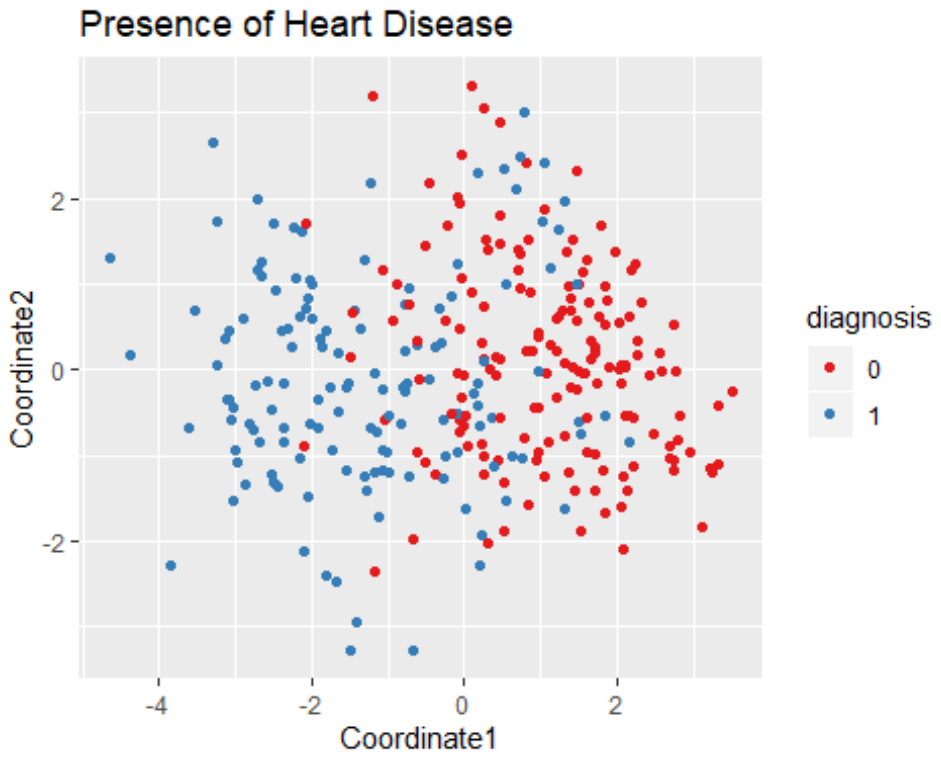
For the clustering, we chose two components based on the "knee" of the scree plot appearing at component 2, as well as the fact that we were only considering two possible outcomes, that is, whether a patient had heart disease or not, so it made sense to include two components in the cluster analysis. We used the K-means and Ward clustering methods with two components because these were the methods that resulted in the highest percentages of explained point variability. We produced a test to compare if the clusters were different between the two methods. The test indicated that the methods were different, so we chose the K-means method over the Ward method based on the graphical representations.

Another way we explored the data was through a correlation matrix with all 14 variables. Most of the relationships were not significantly correlated, but there were two that were. The following graphs represent the relationships between the two pairs of variables: old peak & slope and num & thal. We wanted to incorporate these findings in our model creation.

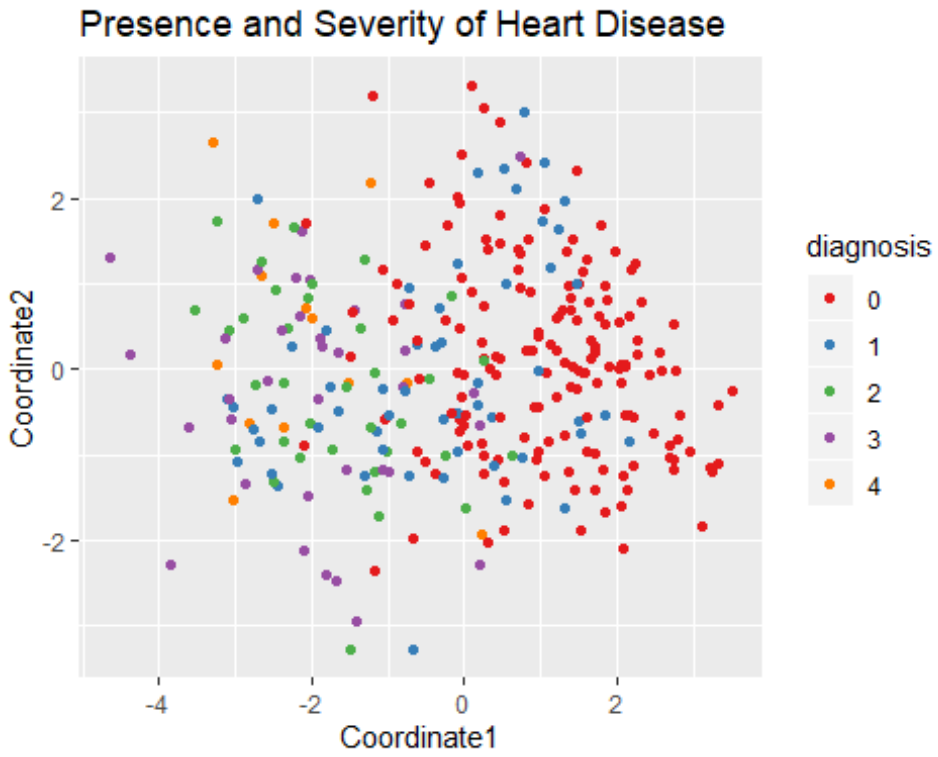
Other Interesting Aspects of the Data

During our data exploration, we came across an interesting observation regarding our diagnoses. Our goal was to see if there was a visual difference among our individual observations. We first created a distance matrix from our observations using all recorded variables, except for the individuals' diagnosis. By not including the diagnosis within our matrix, we are able to see how different each individual is from one another, based solely on their other health measurements. In a way, we're categorizing our individuals by their

diagnosis, without using their diagnosis as direct form of measurement. We then used multidimensional scaling to produce a data set where each row in the data represents an individual observation, and all of their recorded traits and symptoms are now defined by a single x-y coordinate. By plotting these points and coloring them based on their presence of heart disease, or lack of, we can see two distinct groups in the graph below.



Based on this graph, it's fair to assume that there is a distinct difference in the health measurements between patients with and without heart disease, but what about the varying levels of severity in heart disease? To determine if there was a significant difference among the severities of heart disease, we repeated the above process, but instead colored individuals based on their severity, or lack of heart disease. The graph below shows there is still a distinct difference between those without heart disease and those with any level of severity. But, if we just look at patients with heart disease, there doesn't appear to be a significant difference in their health measurements, in just two dimensions.



If we were to look at this same data set and use multidimensional scaling in three or four dimensions, it's possible that we could see a significant difference in the severity of heart disease. For our purpose, we decided to not further investigate or predict the severity of heart disease, and instead just predict the presence or lack of it.

Model Creation and Selection

In order to create a model to efficiently and accurately predict heart disease in a patient, we started with a full model to include all variables then attempted to eliminate insignificant variables while ensuring accuracy. In our first attempt to eliminate variables, we decided to use the clusters we found earlier to decide which variables to keep and which to eliminate. Since the clusters showed which variables behaved similarly, we kept only one variable from each cluster, choosing the one that was the most significant from the full model. That model included ST depression induced by exercise relative to rest, chest pain type, number of major vessels colored by fluoroscopy, resting electrocardiographic results, thal diagnosis, maximum heart rate achieved, and resting blood pressure. To decide between our models, we compared the area under the curve (AUC) for each model; a higher AUC indicates a better model. The second model had a slightly lower AUC compared to the first, but it was close, so it was still accurate. We wanted to see if a smaller model could still give us good predictions, so we tried a different method to eliminate variables. For the third model we created, we chose to go back to the full model, and eliminate any insignificant variables (anything with p-value > 0.05). This left us with sex, chest pain type, slope of peak exercise ST segment, number of major vessels colored by flourosopy, thal diagnosis, and exercise induced angina. The third model gave us a similar AUC as the first two, and through and analysis of variance test, it was determined that the third model was not

significantly different from the full model in terms of accuracy, so we chose the third model going forward since it included the least amount of variables. In addition to looking at the significance of the variables, we wanted to make sure to include the thal diagnosis variable since we saw in the exploratory section that it was highly correlated with the presence of heart disease. Also found during exploratory analysis was the relationship between the old peak and slope variables. Since the two were highly correlated, we wanted to make sure only one of those two was included in the final model to avoid redundancy. As another measure of accuracy for the chosen model, sensitivity (true positive rate) and specificity (true negative rate) were also calculated. After creating our model, we were able to use it to predict if a person has heart disease given their chest pain type, slope of peak exercise ST segment, number of major vessels colored by fluoroscopy, thal diagnosis, and exercise induced angina.

Results and Conclusion

In order to arrive at the best model for predicting whether or not a patient has heart disease, we compared the three general linear models we created. We determined that the reduced model including all significant variables was best. Again, this model had the variables sex, chest pain type, slope of the peak exercise ST segment, number of major vessels colored by flourosopy, thal diagnosis, and exercise induced angina. Any other variables were not deemed insignificant in determining if a patient has heart disease or not. Looking at a ROC curve using the final model created, the threshold of greater than .529 was found for predicting that the patient has heart disease. In terms of measuring accuracy, we used the misclassification rate, sensitivity and specificity. The misclassification rate of our model, using the .529 threshold, was 14.85%, the specificity was 85%, and the sensitivity was 90%. Therefore, with our model, we correctly predict that a patient has heart disease 90% of the time, and predict that a patient does not have heart disease correctly 85% of the time. When using our model to predict diagnosis of heart disease for a specific patient, we were only interested in the variables that made up our model. For example, using our model to predict for a female patient with non-anginal pain, no exercise induced angina, upsloping peak exercise ST segment, 1 major vessel colored by fluoroscopy, and normal thal diagnosis, the model gave a value of 0.1507 for the response variable of heart disease diagnosis. Therefore, using the threshold of greater than .529 as having heart disease, we can conclude that this specific patient does not have heart disease.