Heart Disease: Contributing Factors and Their Relationships

Olivia Bojesen

11/05/2023

# Introduction

Throughout this analysis, we were interested in determining if there are certain factors that are associated with a higher risk of heart disease. Using this information, we wanted to create a model that can accurately predict whether a new patient is at risk for heart disease based on the measurements given.

# Description of the Data

The data initially had 303 patients included, however six observations were removed due to random missing values, so the following analysis was done on the remaining 297 patients. The majority of these 297 patients (53.87%) indicated no heart disease. Those with heart disease were assigned numbers 1 through 4 to indicate increasing levels of heart disease severity and as heart disease severity increased, observations decreased. The variables considered included demographic data such as age and sex, along with information from various medical tests such as chest pain type, resting blood pressure, and maximum heart rate achieved. Eight of the fourteen variables were categorical. Due to this, many of the analyses performed were done on the remaining six numerical explanatory variables.

# Methods

All analysis on this historical medical data was done in R. Exploratory analysis was used to determine associated groups of variables that may exist. We used Bartlett’s test with a significance level of 0.05 to determine that factor analysis was an appropriate method for the data, which was used to find underlying groups of the six numerical variables. We decided which variables were fit for factor analysis by using the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA) to keep variables with an MSA above 0.5.

Using the principal component method of factor analysis, the number of components and the variables kept were chosen based on retaining variables with communalities (h2) greater than 0.5. Extracting three components using all six variables fit these criteria and was determined to be the best way to have as much shared variance among variables as possible without getting rid of too many variables, which was the case when extracting only two components.

K-means clustering was used to determine groups of patients that may exist. The number of clusters, k, was chosen to be the point before the within cluster sum of squares began to level off on the scree plot. The clusters of patients were colored on the principal component plot to visualize how they related to the variables associated with the principal components.

Due to the large number of variables (thirteen), we looked at different ways of reducing dimensions such as principal component analysis (PCA) and non-metric multidimensional scaling (NMDS). To determine the number of components to use for PCA, percent of variance explained was considered. For NMDS, we used a stress level of 10 to represent a good fit. However, for both PCA and NMDS, the understandability of the visualizations produced, and the interpretation of the components were important factors for selecting the number of components used in the plots.

Hotelling’s T-square test was used to test if there was a significant difference between the group of patients with and without heart disease. A significance level of 0.05 was used.

We used linear discriminant analysis (LDA) to determine a model for predicting whether a patient is at risk for heart disease. 75% of the data was assigned to a training set that was used to build the model and this was then tested on the remaining 25% of the data. The accuracy of the model was determined by the following measures: misclassification rate, sensitivity, specificity, and the area under the ROC curve. Although LDA is not intended for categorical data, the model proved to be accurate enough for us to move forward with this method. Weighting of the group with heart disease was increased to produce a sensitivity percentage above 80%.

# Results

Conducting factor analysis on the six numerical variables in the dataset showed associations between the number of major vessels colored by fluoroscopy, ST depression induced by exercise relative to rest, age, and maximum heart rate achieved, with maximum heart rate being negatively related to the former three variables. Using three components explained 68% of variance, so while the model is missing out on some unexplained variance, it gives us an idea of groups formed by the variables as we move into further analysis.

Figures 1 and 2 below show relationships between some of the associated variables mentioned. In figure 2 we can see that more patients with heart disease are older and have a lower maximum heart rate.

A diagram of a heart rate

Description automatically generated

A diagram of heart rate and age

Description automatically generated

Using k-means clustering with k = 4 components gave us insight into the groups the patients formed based on the six numerical variables we had. The principal component plot below (Figure 3) shows the four clusters by color.

A diagram of different colored dots

Description automatically generated

Figure 3 above displays the clusters plotted against the first two components, which account for about 53% of variance. While there is more dimensionality to the data than what is seen here, this plot allows us to easily visualize all six variables in just two dimensions. Component 1 reiterates our findings from the factor analysis completed above, where a higher value for the first component indicates high values for number of major vessels, ST depression, and age, along with a low maximum heart rate. Component 2 is composed of high values of serum cholesterol, resting blood pressure, and maximum heart rate.

Figure 3 also shows whether a patient has heart disease, as seen by the shape of the points. As we can see, those with heart disease tend to be further to the right with respect to component 1. We can also see that cluster 1 has fewer patients with heart disease than the other 3 clusters and is further to the left. Both of these observations would suggest that higher age, more major vessels, higher ST depression measurement and lower maximum heart rate are associated with heart disease.

Unlike PCA, non-metric multidimensional scaling (NMDS) was able to reduce dimensions of all thirteen variables of interest, categorical variables included. We used five components to reduce the dimensions of our original thirteen variables, which gave a stress value of 10.23, indicating a fairly good fit. Figure 4 below plots the patient data and their heart disease status with respect to the first two coordinates.

A diagram of red and blue dots

Description automatically generated

Coordinate 1 has a clearer association with heart disease than coordinate 2. While there is overlap in the middle of the plot, we can see that patients with heart disease have higher coordinate 1 values. This indicates that there are similarities in patients who have heart disease.

We used the Hotelling’s T-square test to determine if there is a statistical difference in any of the variables between the group of patients with and without heart disease. This test resulted in a p-value of 0, showing that there is a significant difference in at least one of the measurements between the two groups. Simultaneous confidence intervals showed that patients with heart disease are older, have higher ST depression measurements, have more colored major vessels, and achieved a lower maximum heart rate than those without heart disease.

With our data, we were able to create a model using linear discriminant analysis that can predict the presence of heart disease for patients whose measurements are known for all 13 explanatory variables. Our model was fit using 75% of the data and was then tested on the remaining 25% of the data to determine accuracy. Table 1 shows different measures of accuracy.

| **Misclassification Rate** | **Sensitivity** | **Specificity** | **Area Under the Curve** |
| --- | --- | --- | --- |
| 13.33 | 77.78 | 94.87 | 87.27 |
| Table 1 |  |  |  |

While the specificity percentage being high (94.87%) is good, meaning few patients are incorrectly diagnosed with heart disease, the sensitivity is low. Since sensitivity measures the percentage of those with heart disease who are correctly identified as having heart disease, we want this percentage to be high. By giving more weight to those with heart disease, we can increase this percentage, however, this will increase the misclassification rate and decrease the specificity. Table 2 below shows the measures of accuracy with this new model.

| **Misclassification Rate** | **Sensitivity** | **Specificity** | **Area Under the Curve** |
| --- | --- | --- | --- |
| 14.67 | 83.33 | 87.18 | 86.3 |
| Table 2 |  |  |  |

Given a patient with the following information:

| **New Patient** |  |
| --- | --- |
| Age | 60 |
| Sex | Female |
| Chest pain type | Non-anginal pain |
| Resting blood pressure | 102 mmHg |
| Cholesterol measurement | 318 mg/dl |
| Fasting blood sugar | Low |
| Resting electrocardiographic results | Normal |
| Maximum heart rate | 106 beats/minute |
| Exercise-induced angina | None |
| ST depression induced by exercise relative to rest | None |
| Peak ST segment | Upsloping |
| Number of colored major vessels | 1 |
| Thal diagnosis | Normal |
| Table 3 |  |

our model was able to predict with 98.9% confidence that heart disease is not present.

# Conclusion

In this analysis we found that a lower maximum heart rate along with higher age, number of colored major vessels, and ST depression are related factors in patients and are all significantly different between patients with and without heart disease. We were able to visualize our data in 2 dimensions using non-metric MDS which showed that there are similarities in the two groups of patients: those who have heart disease and those who don’t. The final model created can correctly predict the presence of heart disease in a patient approximately 85% of the time.