Heart Disease Patient Analysis

Julia Haas

2023-11-03

**Introduction and Methods**

The purpose of this report is to investigate the historical dataset processed.heartdisease.txt that includes 303 patients at risk of heart disease displaying 14 different recorded fields. The data contains a mix of quantitative and categorical variables and has six different instances of missing data. To address the missing data, the variables have been marked as NA and were not included in the different analysis preformed. I chose not to generate predictions for the missing data because those predictions have the possibility of inaccurately representing the patient. Medical researchers were interested in the following questions:

1. **Are there notable associations/relationships between some of the variables? Are there any meaningful groups of variables that exhibit these associations? If so, describe them.**

After creating a correlation matrix of the quantitative variables which include age in years, resting blood pressure, serum cholesterol, maximum heart rate achieved, and ST depression induced by exercise relative to rest, no strong correlations between two variables presented themselves. Serum cholesterol and maximum heart rate achieved appeared completely uncorrelated from one another, and there are weak negative relationships between maximum heart rate achieved and age, and ST depression induced by exercise relative to rest and maximum heart rate achieved. By preforming a Pearson's Chi-squared test on thal and diagnosis of heart disease, there seems to be an association between these variables.

## age trestbps chol thalach oldpeak  
## age 1.00 0.28 0.21 -0.39 0.20  
## trestbps 0.28 1.00 0.13 -0.05 0.19  
## chol 0.21 0.13 1.00 0.00 0.05  
## thalach -0.39 -0.05 0.00 1.00 -0.34  
## oldpeak 0.20 0.19 0.05 -0.34 1.00

Pearson's Chi-squared test

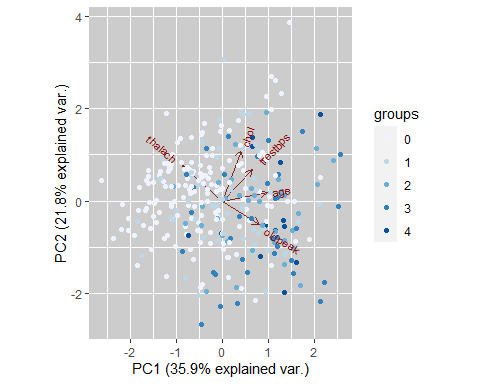
X-squared = 97.994, df = 8, p-value < 2.2e-16

1. **Is there a way to graphically represent the raw data for the 303 patients and draw conclusions about the data set from such a graph? graphs, visualizing the data?**

To graphically represent the raw data for the 303 patients we can perform a principle component analysis. To do this only the quantitative variables may be examined. Because each variable has a different unit of measurement the data must be scaled and a correlation matrix must be used. When considering how many components to include, a scree plot suggests that 2 components may be enough. Following Kaiser’s rule of having a standard deviation of 1, we would again need 2 components. When looking for a variance explained of 70%, we would need 3 components. Overall, I chose to include the first two components. The first component is impacted heavily by age in years, ST depression induced by exercise relative to rest, and maximum heart rate achieved. The second component is impacted heavily by resting blood pressure, serum cholesterol, and maximum heart rate achieved. From the graph we can see that people with more severe levels of heart disease seem to be older in age with low A graph with a line

Description automatically generatedmaximum heart rates, and higher resting blood pressures. In the figure, “groups” represents the diagnosis of heart disease (0: no heart disease; 1-4: increasing levels of heart disease severity).  
## Loadings:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## age 0.568 0.113 0.181 0.518 0.603  
## trestbps 0.382 0.427 -0.732 0.141 -0.341  
## chol 0.241 0.686 0.534 -0.393 -0.179  
## thalach -0.505 0.481 -0.307 -0.171 0.625  
## oldpeak 0.467 -0.321 -0.229 -0.727 0.313

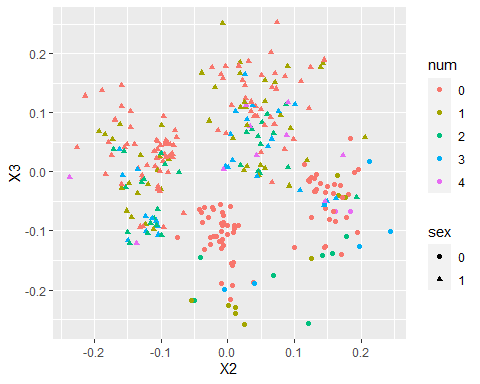
## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 1.3405942 1.0446080 0.9352112 0.8733512 0.68865025  
## Proportion of Variance 0.3594385 0.2182412 0.1749240 0.1525485 0.09484783  
## Cumulative Proportion 0.3594385 0.5776797 0.7526037 0.9051522 1.00000000



1. A graph with numbers and lines

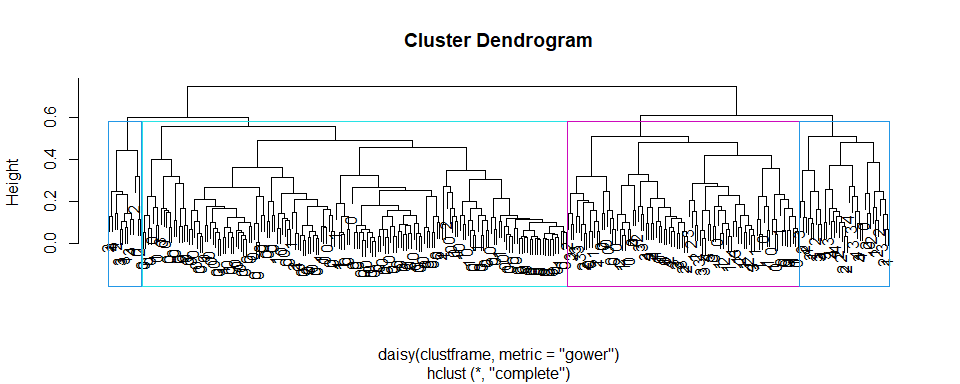
   Description automatically generated**What are the basic underlying groups that the individuals form? Can you plot the data in a small number of dimensions, showing the group separation of the patients?**

To look for possible groups I created a distance matrix of the variables including categorical data. I then used this matrix to perform a nonmetric multidimensional scale with k dimensions equal to 5 to keep the stress under 10%. By graphing different components, we can see that the data does separate itself into groups, but because the MDS components cannot be very well interpreted, these groups are difficult to define. To try and partially understand what kind of groups the graphs visualize I color coded each figure by the diagnosis of heart disease (0: no heart disease; 1-4: increasing levels of heart disease severity, and defined datapoints with the shape circle to represent females and datapoints with the shape triangle to represent males.

A graph with many colored dots

Description automatically generated

I also performed a complete linkage hierarchical clustering of the data and was able to include both categorical and quantitative variables using the Gower metric which creates a dissimilarity matrix of the variables that can be used to create a cluster dendrogram. I used the cutree function to split the clustering into four different groups. From the dendrogram we can see that the first group is relatively small and contains mainly diagnostics of heart disease ranked 2 and 3 in severity. The second group contains almost all non-heart disease patients. The third group seems to contain a range of diagnoses 0 through 3 and the final group is composed of mainly patients with 2, 3, and 4 in heart disease severity.



1. **Are there interesting differences in any of the recorded fields with respect to heart disease diagnosis?**

One interesting difference is that there presents no correlation between maximum heart rate and serum cholesterol in the correlation matrix. This is interesting because in medicine a higher pulse rate is often associated with high cholesterol levels, yet this correlation does not seem to be present in this dataset.

Another interesting difference between variables is shown between maximum heart rate achieved and ST depression induced by exercise relative to rest. The principle component analysis visualizes the two variables as being strongly negatively correlated. This negative association is also represented in the correlation matrix also this negative correlation is not necessarily strong, listed at -0.34.

1. **If the researchers were to investigate a new patient observation that had known measurements for the 13 explanatory variables, could we determine a rule for predicting that patient’s heart disease status (no heart disease vs presence of heart disease)? How accurate could you expect such a rule to be?**

One way of determining a rule for predicting a patient’s heart disease status would be to use linear discriminant analysis. This analysis uses sample information from the historic dataset to fit new patients into a classification. One limitation of this model would be that categorical variables cannot be used in the analysis. Even with that limitation, linear discriminant analysis could still be very useful in predicting heart disease for new patients. One assumption of this analysis is that the variables are multivariate normal and based on three different multivariate normal tests and a multivariate plot, this assumption is not met. To understand how accurate the analysis is without having multivariate normal data, I performed a cross-validation misclassification test which produced a misclassification rate of 0.4356. I decided to continue with the linear discriminant analysis which includes information about age in years, resting blood pressure, serum cholesterol, maximum heart rate achieved, and ST depression induced by exercise relative to rest.

1. **In particular, we have a new patient who is a 60-year-old female. Her symptoms are non-anginal pain, a resting blood pressure of 102 mm Hg, a cholesterol measurement of 318 mg/dl, low fasting blood sugar, normal resting electrocardiographic results, a maximum heart rate of 160 beats/minute, no exercise-induced angina, no ST depression induced by exercise relative to rest, upsloping peak ST segment, only 1 colored major vessel, and normal thal diagnosis. Would you predict this patient to have heart disease? How confident are you in the classification?**

To predict the diagnosis of this patient I used the values of her quantitative variables to create a new data frame. I compared her data to the linear discriminant analysis with the predict function and was able to predict that she does not have heart disease. By looking at the posteriors of this prediction we are 79.58% certain in this classification. This certainty is very likely to be far less than 79.58% because of the lack of multivariate normality in the dataset and the fact that categorical variables like sex, non-anginal pain, low fasting blood sugar, normal resting electrocardiographic results, no exercise-induced angina, upsloping peak st segment, 1 colored major vessel, and normal thal diagnosis are not being taken into consideration with the prediction.

## [1] 0  
## Levels: 0 1 2 3 4

## 0 1 2 3 4  
## 1 0.795752 0.1502112 0.03936124 0.01049122 0.004184328

By creating ratio tables of the categorical variables, we may get a better understanding of how accurate her diagnosis prediction really is. The majority of patients who did not have heart disease had normal resting electroencephalographic results, 57.927%, compared to heart disease patients in which the majority had results showing probable or definite left ventricular hypertrophy by Estes' criteria. The majority of patients who did not have heart disease had low fasting blood sugar, 85.98%. That being said, the majority of heart disease patients also had low fasting blood sugar, 84.17%. The majority of non-heart disease patients had no exercise induced angina, 85.98%, while the majority of heart disease patients had exercise induced angina, 54.67%. The majority of non-heart disease patients had an upsloping peak ST segment, 64.34% unlike heart disease patients who had a majority of flat peak ST segment, 64.47%. In the dataset, 32% of the patients were female and 68% of the patients were male. Of those patients, 82% of patients diagnosed with heart disease were male and 18% were female. The majority of non-heart disease patients had 0 colored vessels, the majority of heart disease patients also had 0 colored vessels, 33.33% with 1 colored vessel a close second at 31.88%. This could be meaningful. In all, the majority of the new patients’ categorical variables point towards a diagnosis of no heart disease. I predict this patient to not have heart disease.