CHARLIE IRMIGER

Researching Significant Variables in Determining Heart Disease

CCS Concepts: • **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: heart disease, datasets, logistic regression, odds ratio

ACM Reference Format:
Charlie Irmiger. 2019. Researching Significant Variables in Determining Heart Disease. *J. ACM* 1, 1, Article 1 (December 2019), 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Heart disease remains as one of the leading causes of death across the country[4], despite the numerous advances in modern health care and research. One of those advancements in research is the explosion of data science and the amount of data available today. Modern technology and computing has allowed us to take incomprehensible amounts of data and analyze it, draw conclusions from it, and to predict future outcomes from past results.

For my research, I found a dataset containing data on heart disease from 303 individual patients, with 14 columns of variables. These columns are:

Age, sex, cp(chest pain experienced), trestbps(resting blood pressure), chol(cholesterol), fbs(fasting blood sugar), restecg(resting electrocardiographic measurement), thalach(maximum heart rate achieved), exang(exercise induced angina), oldpeak(ST depression induced by exercise relative to rest. "ST" relates to positions on the ECG plot), slope(slope of the peak exercise ST segment), ca(number of major vessels that appear on stress test), thal(thallium stress test result), and our independent variable, target(heart disease).

The goal of my research into this dataset is to find variables that have significant influence on the outcome of the target variable. If variables that contribute to heart disease can be accurately identified, then lifestyle changes can be made in order to reduce the chance of heart disease in future patients.

2 LITERARY REVIEW

2.1 Established Causes of Heart Disease

While heart disease is an incredibly complicated and expansive disease to categorize and predict, there are some well established variables that have significant influence on whether or not a subject may have heart disease.

When initially hypothesizing variables that may be of significance, cholesterol, smoking, and age seemed to be factors that I have heard of having significant impact on heart disease. My research into heart disease only strengthened my belief in my hypothesis since I found that the

Author's address: Charlie Irmiger, Marquette University, Milwaukee, Wisconsin, 53233, Charles.Irmiger@marquette.edu.

0004-5411/2019/12-ART1 $15.00
https://doi.org/10.1145/1122445.1122456

main variables to focus on would be cholesterol levels[2](specifically low density lipoprotein (LDL) cholesterol), blood pressure, smoking, and age.[1]

2.2 Diagnosing Heart Disease

There are many methods used to diagnose heart disease, but in general a doctor will perform a physical exam and analyze a blood test before moving on to other tests. My dataset contained variables obtained from numerous different heart tests, such as blood pressure, stress tests, ECG measurements, and thallium stress test result. The thallium stress test is a tracer that is injected into the subject that is then analyzed with a camera or detector to compare segments of the heart.[3]

3 HEART DISEASE DATASET

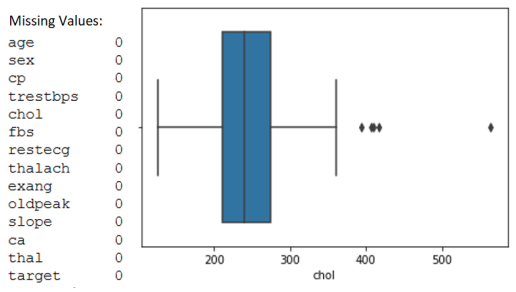
My *Heart Disease UCI* dataset came from "https://www.kaggle.com/ronitf/heart-disease-uci", a service widely used by data scientists to perform data science tasks such as cleaning, regression, and modeling on datasets which can then be shared for others to use and interpret. The *Heart Disease UCI* dataset was originally implemented from the Cleveland Database, which is used by researchers in machine learning topics much like Kaggle. My dataset contained data on 303 individual patients, with both continuous and count variables used to describe information on the patients and the results of various tests to categorize the presence of heart disease. The dataset had 14 column variables and 303 row entries, making this a relatively small dataset.

3.1 Methods

Before I began the process of cleaning and tidying my data, I first wanted to get a general summary of what kind of data I had, the sex of the subjects, and the age range. What I found from my initial analysis was that the average age was 54 years old, and around 32 percent of the subjects were female compared to 68 percent for men.

After my initial analysis on the subjects and their general statistics, I took a deeper look at the rows and columns in order to clean my data. I first looked for missing data by using isna() on each of the columns to count how many data points were missing, which returned 0 meaning that there was a value for every variable in the rows.

The next step was to tidy the data, which was useless since the dataset's column headers were already adequate descriptors of the variables contained in them, and the rows all had valid values and no missing data. This also meant that I was unable to melt the data in any way. However, when looking at the descriptions of the column headers, I found that the "Target" column was used to categorize heart disease, with '0' meaning heart disease and '1' meaning no disease. I found this confusing and switched the values in the target column so that '0' indicated no disease and '1' indicated that the subject had heart disease.



The final step for cleaning and tidying my data was to find any outliers and remove them if they were over three standard deviations from the mean. This was done in order to lower the influence of the point/s on the y-intercept of my regression line. I ended up removing one outlier which had a cholesterol level of 564 mg/ml, compared to the average cholesterol level of 245 mg/ml from the entire dataset.

3.2 Research Questions

After examining my dataset and the descriptions of the columns, I began to look for possible conclusions that could be drawn from the *Heart Disease UCI* dataset. The "target" column was the dependent variable that I chose, since it was a nominal variable that directly categorized the presence of heart disease. This led me to two research questions to explore:

(1. Which column variables have significant influence on the "target" value?

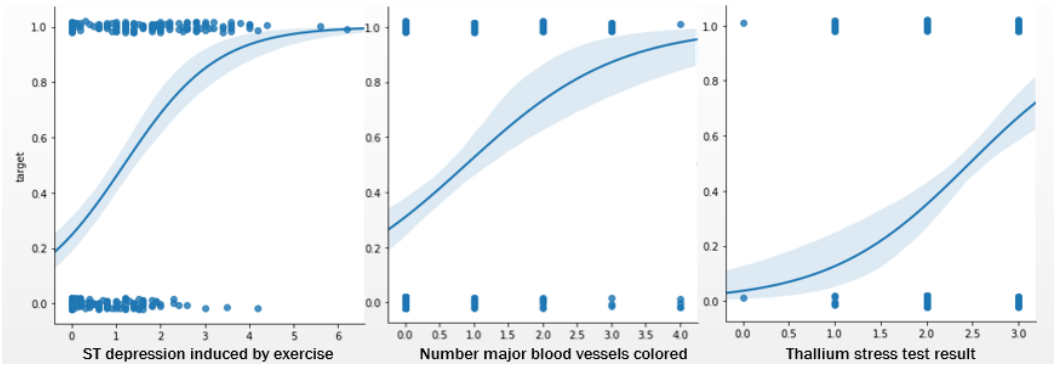
(2. How do the influential column variables compare to each other?

Since the dependent variable in my research question was "target", a nominal variable with only two possible outcomes, I ran logistic regression models on the columns variables in order to determine the probability of each variable having significant influence on the target value being 1. After determining the significant variables, I ran odds ratio tests on them in order to interpret the effectiveness of each predictor.

4 RESULTS

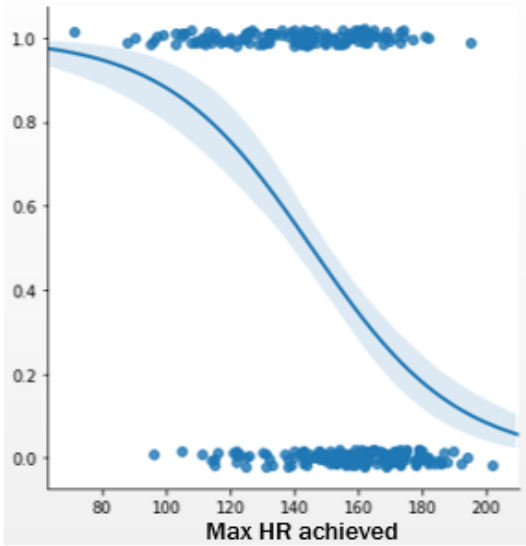
4.1 Initial Visualizations

The first step in determining which variables had significant influence on the target variable was to graph each column variable against the target value using Seaborn's lnplot, which is used for fitting regression models. Graphing each individual column variable against the target value had the effect of breaking down the relationships into small multiples, which is useful for comparisons. I then fit a logistic regression curve to each graph using the built-in method from Seaborn. After each column variable was graphed and the logistic lines were fitted, I was able to easily compare the effect that each variable had on the target value. What was obvious from the graphs was that the column variables that seemed to have the most **positive** influence on the dependent variable were "Oldpeak"(ST depression induced by exercise relative to rest), "Thal"(thallium stress test result), and "Ca"(number of major vessels colored by flourosopy).

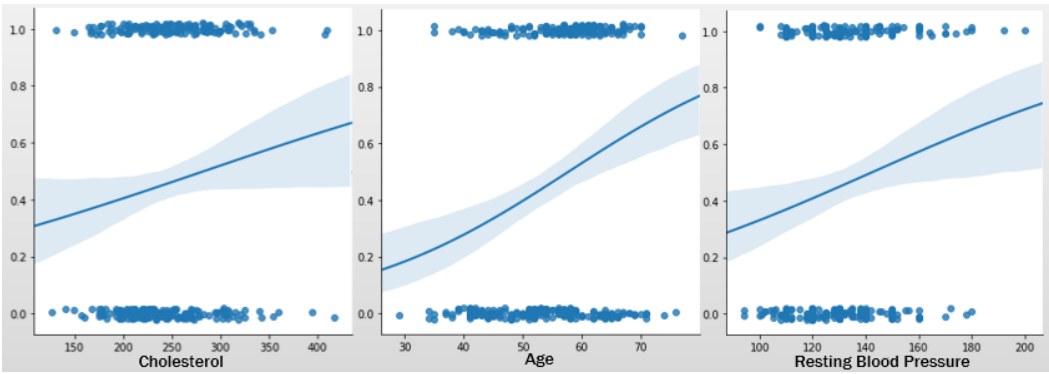


After I had found the most influential **positive** relationships, I found that "Thalach"(maximum heart rate achieved) was the only significant variable with a **negative** relationship.

What I found perhaps most interesting was the fact that age, cholesterol, and resting blood pressure did not seem to have any significant influence on the target value. This was surprising



since my initial research led me to hypothesize that all three would be significant variables in determining heart disease.



4.2 Logistic Regression

The next step in analyzing my data was to perform logistic regression on all variables that seemed to have significant influence on target from my initial visualizations. This was done by grouping all of my newly hypothesized variables into a new dataframe, and then running a logistic regression fit on the dataframe to get the p-values for each variable. I also included age, cholesterol, and resting BP, just to verify that they were in fact not significant.

What I found after running logistic regression was that: ST depression induced by exercise relative to rest, thallium stress test result, number of major vessels colored by flourosopy, and maximum heart rate achieved all had p-values very close to 0, which means that we reject the null hypothesis, and that those four column variables all had significant influence on the target heart disease variable.

For age, cholesterol, and resting blood pressure however, the p-values were all greater than 0.05, meaning that we fail to reject the null. These variables do **not** have significant influence on the target value.

Table 1. P-Value Results

Variable	P-Value
ST Depression	0.000
Thallium Test Result	0.000
Num. Colored Vessels	0.000
Max HR Achieved	0.000
Age	0.318
Cholesterol	0.155
Resting Blood Pressure	0.211

4.3 Odds Ratio

The next step in analyzing my data was to determine what kind of relationship, and how influential the relationship is between the significant variables found in logistic regression tests and the target column.

Table 2. Odds Ratios

Variable	Odds Ratio
ST Depression	1.962437
Thallium Test Result	3.170220
Num. Colored Vessels	2.156946
Max HR Achieved	0.962813
Age	0.983749
Cholesterol	1.004638
Resting Blood Pressure	1.010605

The results from the odds ratio tests *(table 2)* tell us that for ST depression, thallium test result, and number of colored vessels, there is a positive relationship meaning that as those values increase, so does the probability of heart disease. The significant variables also rank from greatest to least in the order of thallium test result, number of colored vessels, and ST depression in terms of the amount of influence they have on the target column.

The odds ratio tests also show us that age, cholesterol, and resting blood pressure all have odds ratio values which are close to 1, meaning that there is no relationship between them and the target value.

Finally, the odds ratio value for max heart rate achieved is less than one, meaning that a negative relationship exists. The higher the max heart rate achieved, the less likely to have heart disease.

5 DISCUSSION

This dataset was interesting to work with and draw conclusions from because the results obtained were very different than what I had hypothesized. I believe that there is greater implication in the fact that variables such as cholesterol and blood pressure *don't* have significant influence over heart disease, as opposed to the discovery of the variables that do. I believe this because these factors are well known in society to be influential on your chance of having heart disease, but when you actually perform logistic regression to back up your hypothesis you find that they are not statistically influential. Misinformation is currently running rampant in our modern society, and it's becoming more important than ever to verify facts for yourself and to not believe everything you hear or see. Commercials for certain foods or drugs that claim to lower cholesterol are selling their products under the assumption that lower cholesterol = lower chance of heart disease, but this can be misleading since the actual data shows us that cholesterol is not a significant variable.

I do however, believe that more research needs to be done on a larger set of data in order to verify the findings from my own dataset. This dataset is relatively small, especially considering the increasing amount of data that is being gathered today. Better research leads to better solutions for problems. The more data we have, the stronger and more accurate our predictions will become, and the more likely that we will be able to predict heart disease in patients that have yet to show symptoms.

6 CONCLUSION

In conclusion, I was able to take my dataset and find column variables that had significant influence on having heart disease. I used logistic regression because my dependent variable was nominal, with 0 meaning no heart disease and 1 meaning heart disease. I fit logistic regression lines onto each of my column variables to determine which had obvious correlation, and then grouped them together to find individual p-values and odds ratios. I had initially hypothesized that variables such as age, cholesterol, and blood pressure would be significant variables, but this dataset showed that the variables ST depression, thallium test result, and number of colored vessels all had positive relationships with the target value, whereas maximum heart rate achieved had a negative relationship with the target value. The values I predicted in my hypothesis were all not significant, which was backed up by the odds ratio tests I performed.

REFERENCES

[1] Shah Ebrahim. 2006. *Multiple risk factor interventions for primary prevention of coronary heart disease.* Retrieved March 2, 2005 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4160097/

[2] William B. Kannel. 1971. *Serum cholesterol, lipoproteins, and the risk of coronary heart disease.* Retrieved March 2, 2005 from https://pdfs.semanticscholar.org/a057/92007d0d0ba19020da6b9edb287778957c0a.pdf

[3] Jeffrey A. Leppo. 1989. *Dipyridamole-thallium imaging: the lazy man's stress test.* Retrieved March 2, 2005 from http://jnm.snmjournals.org/content/30/3/281.long

[4] Stephen Sidney. 2018. *Comparative trends in heart disease, stroke, and all-cause mortality in the United States and a large integrated healthcare delivery system.* Retrieved March 2, 2005 from https://www.sciencedirect.com/science/article/abs/pii/S0002934318302018