Heart Failure/Disease Prediction

Himangshu Raj Bhantana

12/3/2021

Summary

This is a study to see whether certain types of chest symptoms are severe and linked to heart disease, as well as to anticipate general causes of heart disease. As an analytical method, a logistic regression model was used. As a result of this investigation, several chest symptoms variables, as well as other predictors, were statistically examined, leading to the discovery of general causes of heart disease.

Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, killing an estimated 17.9 million people each year and accounting for 31% of all fatalities. Heart attacks or strokes account for four out of every five CVD fatalities, with those under the age of 70 accounting for one-third of these deaths. This study delves deeply into the true causes of heart disease/failure. Some of the specific questions that will be addressed in this study are as follows:

- Determining which sorts of chest symptoms are severe and connected with heart disease, as well as predicting general causes of heart disease.

The summary statistics for the continuous predictor for our heart dataset are shown in the table below.

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Age	918	53.511	9.433	28	47	60	77
RestingBP	918	132.397	18.514	0	120	140	200
Cholesterol	918	198.800	109.384	0	173.2	267	603
FastingBS	918	0.233	0.423	0	0	0	1
MaxHR	918	136.809	25.460	60	120	156	202
Oldpeak	918	0.887	1.067	−3	0	1.5	6
HeartDisease	918	0.553	0.497	0	0	1	1

Data

Before getting into the data analysis for this study, the heart dataset was created by combining multiple datasets that were previously available separately but had not been integrated. This heart dataset combines five previous heart datasets with 11 similar features to provide the largest heart disease dataset available for research purposes to date.

The following are the five datasets used for curation:

- Cleveland has had 303 observations, while Hungary has had 294 observations.
- Switzerland has 123 observations.
- At Long Beach, Virginia, 200 observations were made, with 270 observations in the Stalog (Heart) Data Set.

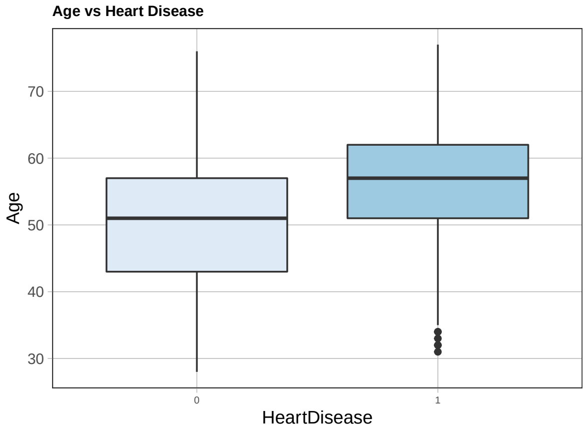
There were 1190 observations in total, 272 of which were duplicated, for a total of 918 observations in the final dataset, with a mix of categorical and continuous variables, which was examined. Because not all variables, such as RestingEC and ExerciseAngina, were applied in this study, since we were primarily interested in heart disease and different forms of chest pain, these variables were also omitted owing to their obviousness in providing the answer straight away before moving on to the next phase; because this was obvious and had a role, we wanted to investigate how the remainder of the variables related to heart disease.

Here is a list of the variables and their descriptions to help us follow along with the rest of the study's process:

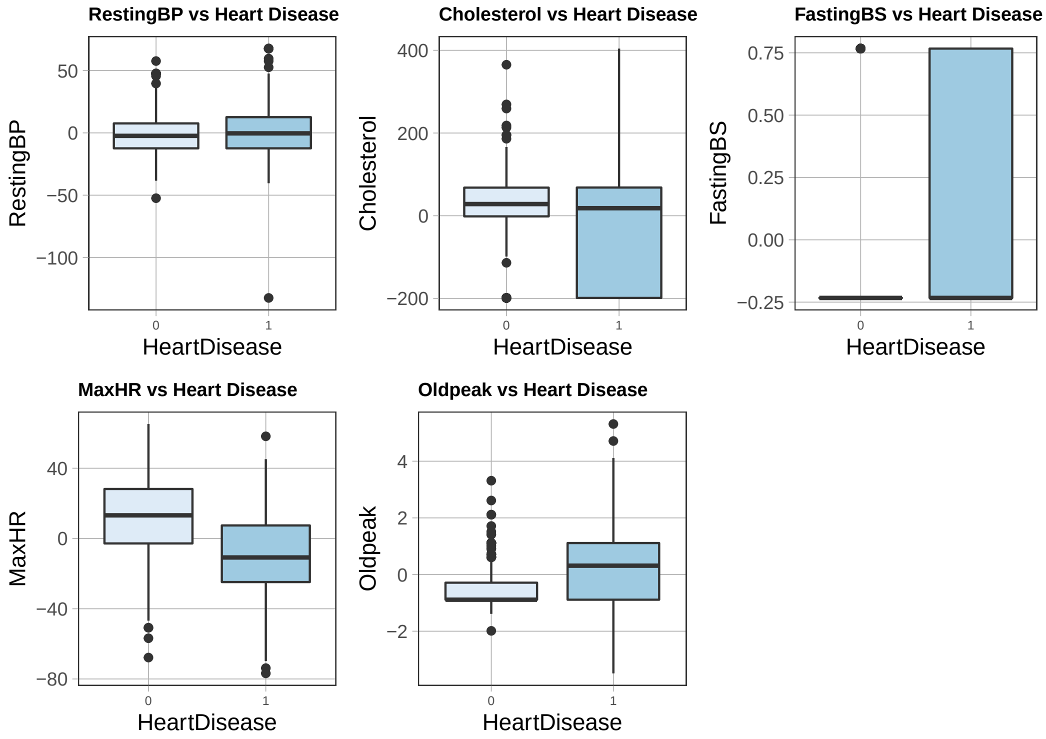
Variable	**Description**
Age	Age of the patient [years]
Sex	Sex of the patient [M: Male, F: Female]
ChestPainType	Chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP	Resting blood pressure [mm Hg]
Cholesterol	Serum cholesterol [mm/dl]
FastingBS	Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
RestingEC	Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
MaxHR	Maximum heart rate achieved [Numeric value between 60 and 202]
ExerciseAngina	Exercise-induced angina [Y: Yes, N: No]
Oldpeak	Oldpeak = ST [Numeric value measured in depression]
ST_Slope	The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
HeartDisease	Output class [1: heart disease, 0: Normal]

Beginning with exploratory data analysis on the updated heart dataset to establish the model and predictors. A factor has been applied to our category variable. The continuous variable was also mean-centered to avoid multicollinearity. Our model is shown in the form of a box plot for the continuous predictor in contrast to our categorical response variable, as well as a table with a CHI-test for comparing both of our categorical variables. Performing this exploratory data analysis allows us to determine if the factors are independent or not. One of the most intriguing EDAs I discovered was the Heart Disease VS Age EDA, which illustrates that as individuals age, they are more likely to get heart disease, which can lead to heart failure.

0 1 410 508



The box plots below indicate the interaction between the continuous variable and the categorical response variable, which is shown to be independent in the majority of cases:



CHI Test

As we work with tables and the CHI test for association for factor variables, we can see whether or not the factor variables are significant. After completing the CHI test for factor variables, it was determined that the response variable heartdisease association with ST_Slope and sex is statistically significant and independent because the P value is less than 5%.

Model

Before fitting the model, we must determine if the predictors are random or not, and if not, a transformation must be applied. We use a binned plot to verify the randomness of continuous predictors. The transformation was unnecessary because the continuous predictors were random and no strange patterns were created. Beginning by fitting the first logistic regression model, which includes all categorical and factor variables, as well as significant correlations and other intriguing interactions, just to see if they influence the model or not.

Because we needed significant predictors to provide the best estimation for our query, I picked the Stepwise AIC model, which generated a new logistic regression model with significant predictors. After experimenting with several selection methods, such as forward and stepwise, which provided the same important predictors, I decided to stick with the AIC Model. Although stepwise regression with AIC advised that I add age as a predictor in my final model, it is negligible at p 0.05 after correcting for the other covariates. As a result, we may infer that the addition of the other factors has eliminated the predictive impact of age.

The mathematical regression equation for the final model is:

$$\begin{aligned}\log\left[\frac{P(\text{HeartDisease}=1)}{1-P(\text{HeartDisease}=1)}\right] &= \alpha + \beta_1(\text{Sex}_{\text{M}}) + \beta_2(\text{ChestPainType}_{\text{ATA}}) + \beta_3(\text{ChestPainType}_{\text{NAP}}) + \\ &\beta_4(\text{ChestPainType}_{\text{TA}}) + \beta_5(\text{ST\_Slope}_{\text{Flat}}) + \beta_6(\text{ST\_Slope}_{\text{Up}}) + \beta_7(\text{Cholesterol}) + \\ &\beta_8(\text{FastingBS}) + \beta_9(\text{MaxHR}) + \beta_{10}(\text{Oldpeak}) + \beta_{11}(\text{Cholesterol}\times\text{FastingBS}) + \\ &\beta_{12}(\text{Cholesterol}\times\text{MaxHR}) + \beta_{13}(\text{FastingBS}\times\text{MaxHR})\end{aligned} \tag{1}$$

The following is a summary of our model selection for our new final logistic regression model:

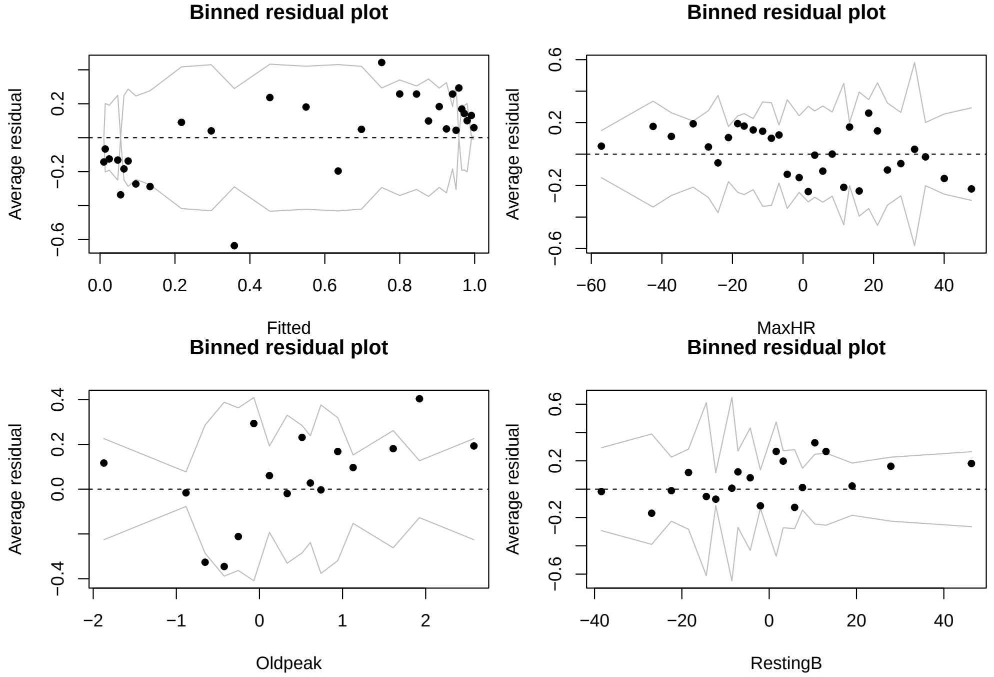
We have yet to determine whether or not there is multicollinearity. When we test the VIP of our final model for multicollinearity, we see that the value for each variable does not exceed the value of 10. This eliminates the need to be worried about multicollinearity.

Table 3: Results

	Dependent variable:
	HeartDisease
SexM	1.60^{***} (0.28)
ChestPainTypeATA	-2.06^{***} (0.32)
ChestPainTypeNAP	-1.81^{***} (0.26)
ChestPainTypeTA	-1.75^{***} (0.43)
ST_SlopeFlat	1.47^{***} (0.43)
ST_SlopeUp	-1.15^{**} (0.45)
Cholesterol	-0.003^{***} (0.001)
FastingBS	1.23^{***} (0.29)
MaxHR	-0.01^{***} (0.005)
Oldpeak	0.51^{***} (0.12)
Cholesterol:FastingBS	-0.01^{***} (0.002)
Cholesterol:MaxHR	0.0001^{*} (0.0000)
FastingBS:MaxHR	-0.01 (0.01)
Constant	-0.36 (0.47)
Observations	918
Log Likelihood	-298.48
Akaike Inf. Crit.	624.96
Note:	*p<0.1; **p<0.05; ***p<0.01

Model Assessment

To evaluate our model, we first generate a binned residual plot of projected probabilities vs. residuals. We can observe that practically all of the observations are inside our confidence interval bands and are dispersed quite randomly. Fitted residuals have points randomly spread about 0 on the Y axis, and there is no evident linear trend, however outliers occur. There is also some clustering on the X axis. While this does not contradict our assumptions, it may imply that the model is lacking certain predictors.

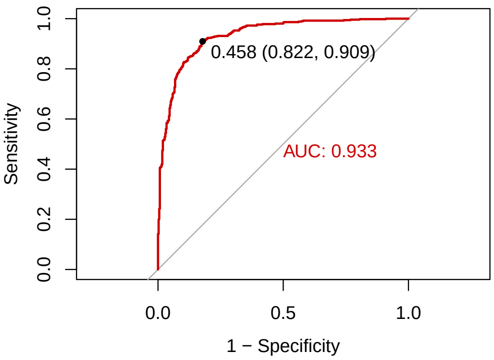


We also plot residuals against each continuous predictor to determine if there is any pattern that our model has missed. When we may conclude that there are no trends when plotted versus residuals

Model Validation

We may now go to model validation. To validate the model, the confusion matrix will be utilized. The accuracy of the model is 85.8 percent, the sensitivity is 86.4 percent, and the specificity is 85.12 percent. The ROC curve shows that using a probability threshold of 0.5 helps us maximize our model's predictions. These numbers appear to be correct, even if it looks like they haven't been totally balanced out, which is impossible. This might be better, but considering the type of data we're working with, our current AUC is adequate for the time being.

The ROC curve is depicted here, and its area is 0.458, illustrating how well our model predicts.



Taking the coefficients for the significant predictors and holding everything else constant yields the following results on average:

- For every unit increase in OldPeak over its mean, the risk of heart disease increases by 1.68 times.
- FastingBS increases the risk of heart disease by 3.4 times for every unit over the mean.
- Every unit increase in MaxHR above the mean increases the risk of heart disease by 0.99 times.
- If all other variables stay constant and the patient has an average cholesterol level, the risk of heart disease increases by 0.99 times.
- Patients with chestpaintypeATA are 0.13 times more likely to have cardiac disease than patients with chestpaintypeASY.
- Patients with chestpaintypeNAP are 0.16 times more likely to have cardiac disease than patients with chestpaintypeASY.
- Patients with chestpaintypeTA are 0.18 times more likely to have cardiac disease than patients with chestpaintypeASY.
- Patients with ST SlopeFlat are 4.4 times more likely to have heart disease than those with ST SlopeDown.
- Patients with ST SlopeFlat are 4.4 times more likely to have heart disease than those with ST SlopeDown.
- Patients with ST SlopeUp are 0.32 times more likely to have heart disease than those with ST SlopeDown.
- Patients with ST SlopeUp are 0.32 times more likely to have heart disease than those with ST SlopeDown.
- Males are five times more likely to have heart disease than females.

Conclusion

To summarize our study's questions, we can draw the conclusion that patients with chest pains such as chestpaintypeATA, chestpaintypeNAP, and chestpaintypeTA have strongly connected with heart disease, which leads to heart failure; additional factors such as whether patients have an oldpeak, fasting blood sugar, cholesterol, and ST Slope Flat and Up are all common factors of heart disease. I'd like to draw our attention to the fact that males are more likely than females to suffer from heart disease or failure.

Limitations

I noted that the data collection for this type of study is insufficient because logistic regression models require a large data set to provide significant results. Despite the fact that multiple merged data sets were used, I feel that having more data would be more effective in predicting heart failure. I also feel that there are other factors that can alter the prediction and make the study more interesting, such as patients who smoke or do not smoke, genetic inheritance, and many others.