Project Paper 2

Alex Ojemann

2023-04-30

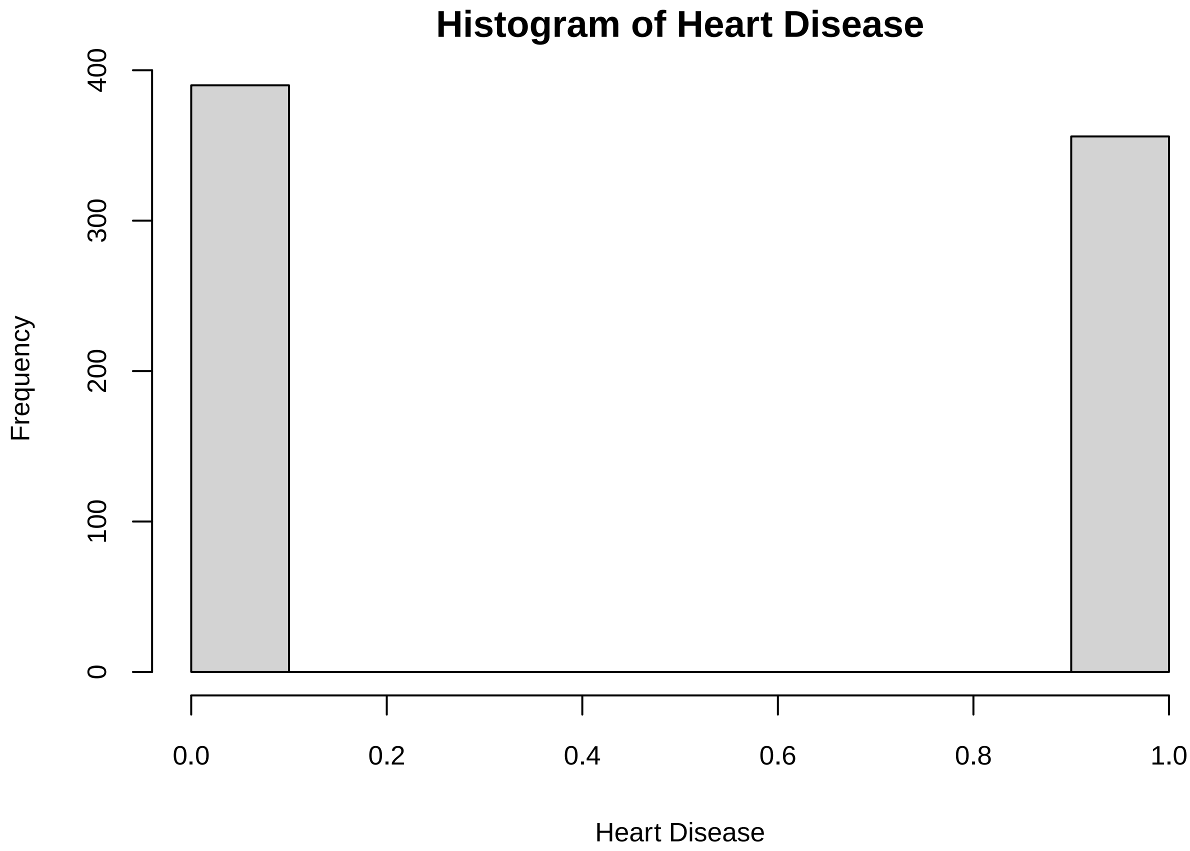
Introduction

My project explores the prediction of heart disease based on some common bio indicators. This data is collected from five different hospitals and contains relevant indicators of cardiovascular disease. It's an interesting problem space because cardiovascular disease is the #1 cause of death globally and it is often said to be significantly related to factors that we can easily measure like blood pressure and cholesterol so it seems like an excellent application for regression. In this paper I will be doing some exploratory data analysis and developing and analyzing my regression model.

Exploratory Data Analysis

The response variable is whether the person in question has cardiovascular disease, a categorical variable represented by a 0 for those that don't have cardiovascular disease and a 1 for those that do. The predictors I will use are age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise induced angina, oldpeak and ST slope.

Response Variable



My response variable is binary, so there can't be any outliers. We can see that there are slightly more patients that don't have heart disease than patients that do. Before significant outliers in some of the predictor variables were removed in Project Paper 1, there were more instances of heart disease than not.

This distribution would not occur if it was sampled from the general population because there are nearly as many people with heart disease as there are people without heart disease in this sample while the proportion of people with heart disease in the general population is much smaller, estimated at 7.2% in the United States. This reflects that this data is not representing the general population of people living near these hospitals, but rather the population of patients at the hospitals, specifically those who are at great enough risk of heart disease to have their biometrics in this data sets measured.

We are 95% confident that the true proportion of patients from the hospitals in this data set that have heart disease is between 0.4413694 and 0.5130542.

Predictor Variables and Multicollinearity

The predictor variables I will include in model selection from the data set are as follows:

Age: The age of the patient in years.

Sex: The sex of the patient.

Chest Pain Type: The type of chest pain. The possible values are typical angina, atypical angina, non-anginal pain, or asymptomatic.

Resting Blood Pressure: The patient's resting blood pressure in mm Hg.

Cholesterol: The serum cholesterol of the patient in mm/dl.

Fasting Blood Sugar: A binary representation of the fasting blood sugar of the patient. 1 if greater than 120 mg/dl, 0 otherwise.

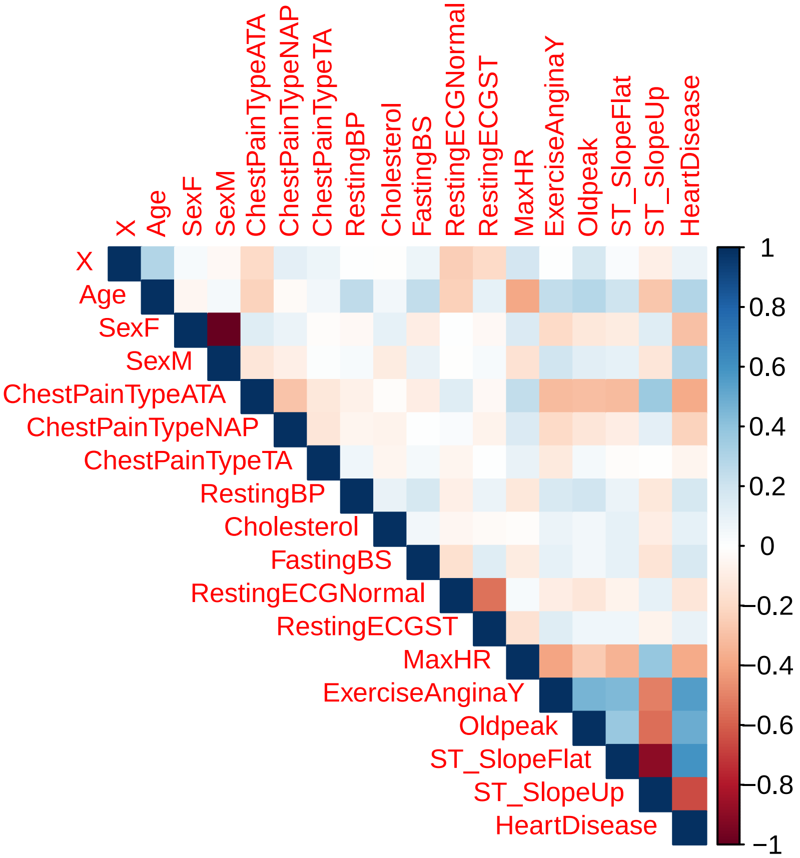
Resting ECG: The patient's resting electrocardiogram (ECG) results. A resting ECG is a non-invasive test that records the electrical activity of the heart while the patient is at rest. The possible values are normal, having ST-T wave abnormality (ST), or showing probable or definite left ventricular hypertrophy (LVH).

Maximum Heart Rate: The maximum heart rate achieved by the patient.

Exercise Induced Angina: Whether the patient has exercise-induced angina, which is chest pain that occurs during physical activity caused by a temporary reduction in blood flow to the heart due to narrowed or blocked coronary arteries.

Oldpeak: A measurement of the amplitude of Q waves in an ECG. This is measured in "depression," or mm below the baseline level.

ST_Slope: A categorical estinate of the slope of the patient's peak exercise ST segment, which is the segment of the waveform between the end of the S wave and the beginning of the T wave in an ECG. The possible values are downsloping, flat, or upsloping.



None of the predictor variables have a correlation > 0.5 (moderate) other than one hot encoded variables within the same category (ST_SlopeFlat and ST_SlopeUp for example).

Most of the relatively high correlation values are between ExerciseAngina, OldPeak, and ST_Slope. These variables may be more advanced and associated with one another, however their correlations are all less than 0.5 thus it's not worth holding them out of feature selection. One of the few pairs of features outside these three that have a correlation magnitude above 0.35 is age and maximum heart rate, which corroborates domain knowledge, but once again the correlation isn't strong enough to assume they're redundant.

```
##                    GVIF Df GVIF^(1/(2*Df))
## X              1.585242  1        1.259064
## Age            1.489559  1        1.220475
## Sex            1.179507  1        1.086051
## ChestPainType  1.293368  3        1.043807
## RestingBP      1.120372  1        1.058476
## Cholesterol    1.071175  1        1.034976
## FastingBS      1.111184  1        1.054127
## RestingECG     1.427059  2        1.092976
## MaxHR          1.537341  1        1.239896
## ExerciseAngina 1.211608  1        1.100731
## Oldpeak        1.444111  1        1.201712
## ST_Slope       1.633810  2        1.130578
```

In addition, none of the predictor variables have a VIF over 5, with the largest being 1.63, so no predictor variables will be removed from feature selection.

Model Development

The features for the model will be selected using best subset selection. Unlike forward and backward stepwise feature selection, which add or remove elements until AIC cannot be lowered any further, best subset selection won’t fall into a local minimum of AIC and will always find the model with the lowest possible AIC because it tries every possible model. This is very computationally expensive because it has to evaluate many more models than forward or backward stepwise selection, but it can be used here because only one model is being created.

```
## Morgan-Tatar search since family is non-gaussian.
```

```
##
## Call:
## glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6810  -0.3930  -0.1140   0.4408   2.9393
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -5.114568   1.247240  -4.101 4.12e-05 ***
## Age               0.033988   0.013335   2.549 0.010809 *
## SexM              1.835024   0.305064   6.015 1.80e-09 ***
## ChestPainTypeATA -1.707323   0.348468  -4.900 9.61e-07 ***
## ChestPainTypeNAP -1.583821   0.295827  -5.354 8.61e-08 ***
## ChestPainTypeTA  -1.609250   0.468669  -3.434 0.000595 ***
## RestingBP         0.013504   0.007128   1.894 0.058163 .
## ExerciseAnginaY   0.888317   0.262320   3.386 0.000708 ***
## Oldpeak           0.403000   0.139741   2.884 0.003928 **
## ST_SlopeFlat      1.255385   0.514898   2.438 0.014764 *
## ST_SlopeUp       -1.292281   0.555841  -2.325 0.020077 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1032.63  on 745  degrees of freedom
## Residual deviance:  487.12  on 735  degrees of freedom
## AIC: 509.12
##
## Number of Fisher Scoring iterations: 6
```

Real World Implications of Estimates

The estimates in the model above represent the coefficients of each predictor variable with the log odds of the heart disease variable. They suggest that:

A higher age results in a slightly increased risk of heart disease, as expected.

A higher resting blood pressure results in a slightly increased risk of heart disease, as expected.

Having Exercise Induced Angina results in an increased risk of heart disease.

A higher oldpeak value results in an increased risk of heart disease.

Males have a higher risk of heart disease than females.

The chest pain types typical angina, atypical angina, and non-anginal pain have very similar associated risks of heart disease but asymptomatic chest pain is associated with significantly higher risk.

A flat ST slope is associated with a higher risk of heart disease than a downward slope, however, an upward slope is associated with lower risk than both. This means that there is no continuous trend in heart disease risk as ST slope increases, rather, it peaks at a flat ST slope.

The only feature that we do not have evidence to say is nonzero at the 0.05 significance level is resting blood pressure, however, we do have evidence to say it's nonzero at the 0.1 significance level so it will not be removed.

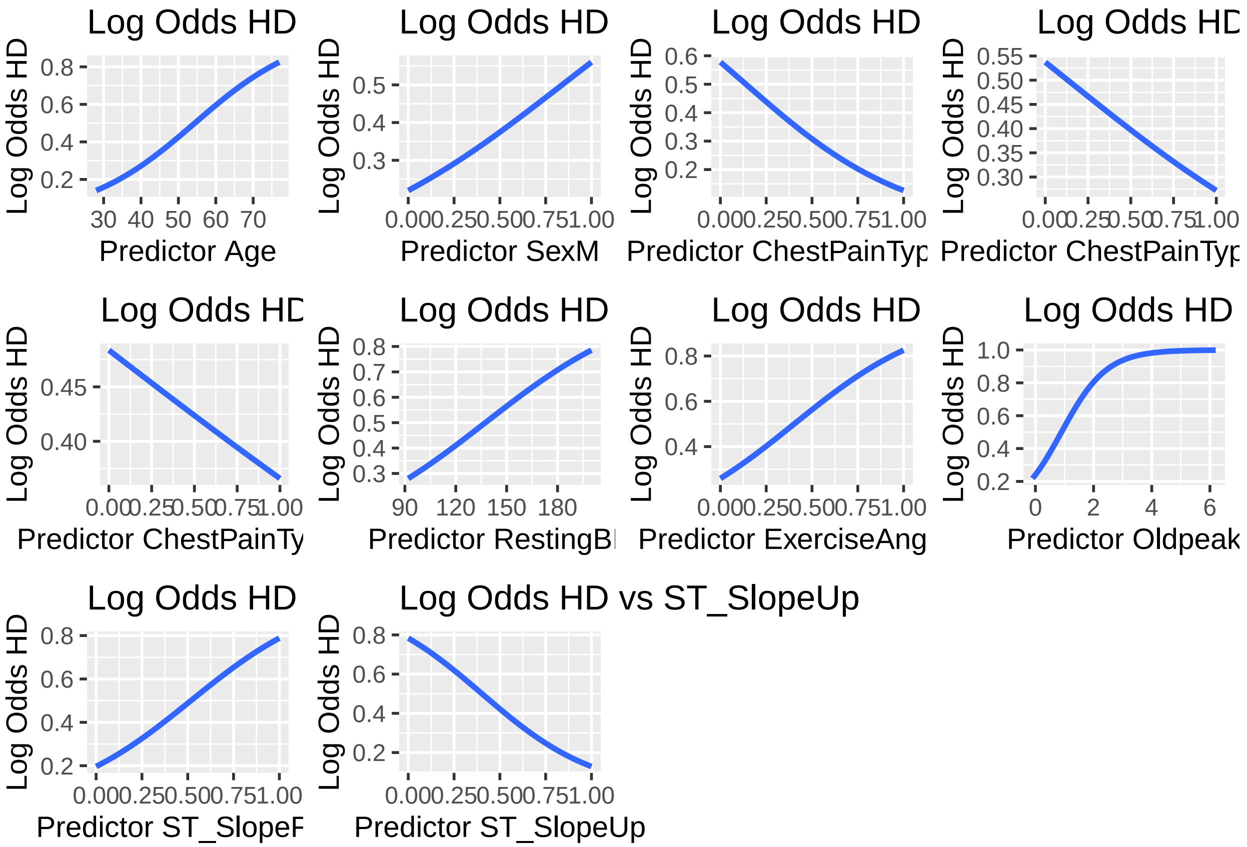
Assumptions

Since our interest is prediction rather than inference, there are just two assumptions that need to be met.

The first is that all of the predictor variables are linearly correlated with the log odds of the response variable. This can be demonstrated visually using the following scatterplots:

```
## Warning: ‘aes_string()‘ was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation ideoms with ‘aes()‘
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



We can see very close to linear trendlines for each of the variables so the linearity assumption is met.

The other assumption that must be met is independence of observations. We would not expect whether one patient in any of the hospitals from the data set has heart disease to affect whether another patient has it, so we expect this assumption to be met. One potential way in which this isn't the case is if relatives go to the same hospital and their shared genetics makes it more likely that they have the same medical conditions, but we do not expect this to occur at a large scale.

Model Analysis

Since our interest is prediction rather than inference, we will evaluate the accuracy of the model using k fold cross validation, specifically using 10 folds as that has been deemed an optimal number of folds by some machine learning experts.

The result is an average accuracy of 0.866 across the 10 folds. A similar application of logistic regression to predict heart disease in the Journal of Pharmaceutical Negative Results reported an accuracy of 0.8868, a similar value.

Real World Examples and Applications

One example of the use of this predictive model is to predict the possibility of heart disease for someone who believes they may be at risk. Not only can the model classify whether they have it with over 86% accuracy, it can give an associated probability. This is useful because a person with 90% risk of heart disease may need to be more urgent in seeking treatment than someone with a 55% risk, even though both would be classified as having heart disease by the model. On the whole, with the great expense that medical care is for many, this can be a useful tool in determining whether treatment and/or preventative measures are necessary.

In addition, this moodel can be useful for understanding the meaning of each of the predictor variables. It is relatively common knowledge that a higher age and resting blood pressure result in an increased risk of heart disease, but those that aren't experts in the field may not understand the meaning of oldpeak or ST slope and this provides numerical evidence for their effect on heart disease.

Conclusion

In this paper I've thoroughly analyzed the data, built a predictive model using exhaustive feature selection techniques to ensure the best possible model in terms of AIC, and computed its accuracy using cross validation. An accuracy of 0.866 exceeded my expectations and is very close to the accuracy of a model from a significant research paper. Other than that none of the results surprised me much as the commonly known predictor variables all had slope estimates that made sense in context.

One thing I would do if I had more time for this project is to create and explore a contingency table to dive further into evaluating the performance of the model than just reporting the accuracy from cross validation. This would be ideal because the two types of misclassifications in the context of the problem are not of equivalent consequence. If the model says the patient has heart disease they may spend money on medical care that they didn't need to, but if the model says they don't have it and they really do, they could have serious medical consequences and die, which is obviously worse, so we may want to accept more of the former to reduce the rate of the latter.