**Analysis of Factors Contributing to Heart Disease**

Olivia Grill

American Public University System

MATH499: Senior Seminar

Dr. Rasoul Behboudi

December 18th 2021

**Abstract**

The data being used for this project is the Cleveland data set from UCI, and the project is completed in R. A combination of exploratory analysis, checking assumptions of logistic regression, and predictive methods were used to produce two different logistic regression models to predict the diagnosis of coronary artery disease. The data provides a dichotomous response variable and 15 predictors variables. Upon completing exploratory analysis and checking logistic regression assumption, I used different predictive approaches in order to create two models. For the first model, I used the generalized linear model (glm) function in R and utilized stepwise processes (16, 1). For the second approach, I used the glm function as well, but this time I introduced the likelihood ratio test with a stepwise process in order to obtain a model that predicts the odds that a person’s coronary artery disease results in a ‘1’, or positive diagnosis. The second approach was added to the study due to the first approach only examining the binary response variable and continuous independent variables. Once both approaches were complete a confusion matrix was calculated to showcase the accuracy of each model. Due to univariate outliers being present for some numeric variables, each approach discussed above was done with outliers and again with outliers removed to examine the potential influence.

**Introduction**

According to the CDC, heart disease has been the leading cause of death in the US since 1921, and still is as of 2020 (26, 4). A specific type of heart disease, Coronary Artery Disease (CAD), is the most common and kills roughly 360,900 people a year (18). CAD is the type of heart disease discussed in this study. There are certain factors that may put people at risk for developing CAD such as high blood pressure, high cholesterol, sex, old age, smoking, diabetes, and being overweight (17). Examining these variables, as well as others, is there a combination of them using a logistic regression model that accurately predicts the probability of someone having CAD? How do the results of the models derived in this study compare to the risk factors and symptoms explained by the CDC? Blood pressure, sex, age and cholesterol are measured directly in this study and this information was considered in the model building process.

There are 11 categorical variables and five numerical variables included in this data set. Of the 11 categorical variables, five of them are binary. Originally there were only nine categorical variables, but I created columns for ‘rbp’ and ‘chol’ that indicate different levels of measure. For ‘rbp’ the additional variable I created is ‘rbpcode’. Using the CDC’s guide, I categorized each ‘rbp’ value into levels of normal, prehypertension, or hypertension (19). For ‘chol’, I created a variable, ‘normchol’, that categorizes each cholesterol level into normal or high based on the threshold value given by MedlinePlus (10). Cholesterol and blood pressure are thought to be some of the main factors in developing CAD, so different levels of each variables may be of significance in this study (17). In R, variables have an “x.” in front of them, but for simplicity, I’m only going to refer to the actual variable name. For example, the variable “age” in R looks like “x.age”, but I’m simply going to refer to this variable as “age”, just know they are referencing the same thing. The variables in the study are given below.

|  |  |  |
| --- | --- | --- |
| Variable Name | *Type* | Description |
| age | *numerical* | Ages of people, ranges from 29 to 77 |
| sex | *binary* | Male = 1, Female = 0 |
| cp | *categorical* | Type of chest pain  0-typical angina  1-atypical angina  2-non-anginal  3-asymptomatic |
| rbp | *numerical* | resting blood pressure when admitted to hospital (systolic only) |
| rbpcode | *categorical* | categorized into normal, pre-hypertension, and hypertension |
| chol | *numerical* | cholesterol in mg/dl |
| normchol | *binary* | categorized cholesterol into normal and high  1-high  0-normal |
| fbs | *binary* | fasting blood sugar, testing whether it was above 120 mg/dl  1-yes/true  0-no/false |
| restecg | *categorical* | resting electrocardiographic results  0- normal  1- ST-T wave abnormality  2-probable or definite left ventricular hypertrophy |
| mhr | *numerical* | maximum heart rate achieved during thallium stress test. |
| exang | *binary* | exercise induced angina (chest pain)  1-yes  0-no |
| oldpeak | *numerical* | ST depression induced by exercise relative to rest |
| slope | *categorical* | Slope of the peak exercise ST segment  0-upsloping/positive  1-flat  2-downsloping/negative |
| mv | *categorical* | number of major vessels blocked when colored by fluoroscopy.  0 vessels show up  1 vessels show up  2 vessels show up  3 vessels show up |
| thal | *categorical* | During thallium stress test  0-normal  1-fixed defect  2-reversible defect |
| cad | *binary* | whether coronary artery disease is present  0-not present  1-present |

**Background**

Multiple of the variables given above are unclear in some form. I want to spend the first portion of this section to discuss them before going into calculations. The first thing I did was rename variables whose name was either too long or non-descriptive of what it was measuring. The renamed variables are given below along with their original name.

|  |  |
| --- | --- |
| **Previous Name** | ***New Name*** |
| trestbps | rbp |
| ca | mv |
| thalach | mhr |
| condition | cad |

Variables I want to discuss further in order to understand their potential contribution to CAD are presented below.

* **thal**- this variable refers to the type of defect detected during a thallium stress test. This test monitors blood flow during rest and exercise (24).
* **mv-** mv stands for major vessels, and this variable measures the major vessels that are visible when using fluoroscopy (13).
* **slope**- Slope refers to a section on an electrocardiogram measurement of the ST segment. Upsloping or down sloping ST segment indicate some form of damage or injury. Flat slop indicates normality (12, 20).
* **oldpeak**- This is another measurement referring to an ECG and is similar to the variable slope. It examines the peak of the line that the slope is based on, ST. It examines ST depression by conducting an exercise test.
* **restecg**- This measure the electrocardiographic results using Estes Criteria. Basically it examines how the heart beats and classifies it as normal or abnormal (3).
* **mhr**- the max heart rate achieved during an exercise test. According to CDC your max heart rate can be calculated by taking 220 and subtracting your age. Then for moderate-intensity physical activity you should be between 64% and 76% of your max heart rate according to age. For high-intensity exercise you should expect the number to be 77% to 93% of your max heart rate (23).
* **exang**- Angina refers to a chest pain that is felt by people who have a lack of oxygenated blood flowing to their heart. This particular variable looks at angina that is caused from exercise (6).

Now that variables have been discussed in further detail, let’s move on to procedures. The first step in the model building process in exploratory analysis. I did this in three main steps, first univariate, then bivariate, and lastly multivariate analysis. There are also five major assumptions that need to be checked for logistic regression (21).

1. The dependent variable is binary

2. Observations are independent of each other

3. No multicollinearity

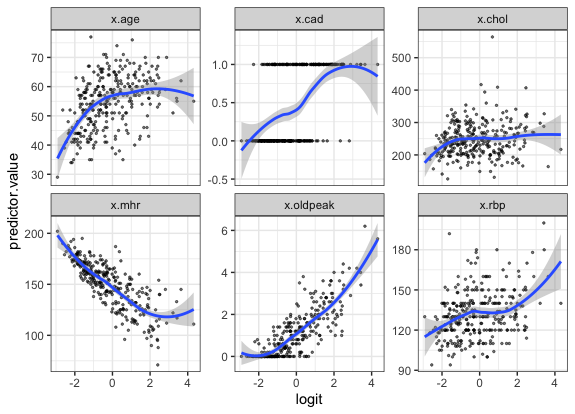
4. No extreme outliers

5. Linear relationship between independent variables and the logit transform of the dependent variable.

**Methodology**

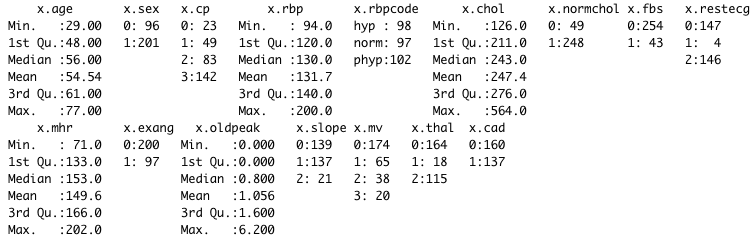
*Exploratory Analysis - Univariate*

First, I wanted to make sure that the dependent variable is binary. Since ‘1’ indicates a person has CAD and ‘0’ indicates that a person doesn’t have CAD, there are only two outcomes for the dependent variables, which confirms it is binary. Since the data wasn’t obtained from repeated measure or matched data, I can also conclude that the observations are independent of one another. I then wanted to do was check the linearity between the log odds of the dependent variable and continuous independent variables (21). The plots for checking linearity are provided below.



Ignoring the ‘cad’ plot, it’s shown that all the independent variables are linearly related to the log odds of the dependent variable. Multicollinearity and outliers are all discussed below, in no particular order.

Next, I made bar plots of categorical variables, histograms of numerical variables, as well as 6 number summaries for all 16 variables. This is an important step as it gives important insight about variables. For example, it lets me examine what proportion of this study was diagnosed with CAD and provides me with an idea of information that I might want to explore in bivariate analysis. Another useful discovery from univariate analysis was looking at the average values of cholesterol and blood pressure. As previously discussed, these are thought to be contributors to developing CAD, and the averages deviate far from what is considered normal (19,10). The graphs and 6 number summaries from univariate analysis is given below.



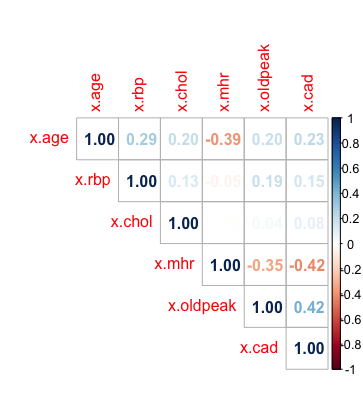
|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

*Exploratory Analysis-Bivariate*

For bivariate analysis, I started out by making a correlation plot and obtaining correlation coefficients of only the numeric variables to examine the associations between variables. R by default calculates Pearson correlation coefficients, but for this data set since the dependent variable is categorical/binary, I opted for the Kendall rank correlation (27). The correlation coefficients range from 1 to -1 indicating a perfectly positive and perfectly negative relationship, respectively. Thresholds vary depending on the area of study, but for this, a table from Boston University School of Public Health was used (25).

|  |  |
| --- | --- |
| **Correlation Coefficient (r)** | **Description (Rough Guideline )** |
| 1.0 (+ or -) | Perfect positive association |
| 0.8 to 1.0 (+ or -) | Very strong association |
| 0.6 to 0.8(+ or -) | Strong association |
| 0.4 to 0.6(+ or -) | Moderate association |
| 0.2 to 0.4(+ or -) | Weak association |
| 0.0 to +0.2(+ or -) | Very weak or no association |

The correlation plot for the variables used in this study is shown below.



Associations between all the continuous variables (ignoring ‘cad’ column) is below 0.4. This indicates that all these variables can be included in the model, as they aren’t highly correlated with one another. None of the variance inflation factor (VIF) values for all the numeric variables aren’t above 1.3 which indicates multicollinearity isn’t present.

The next step in bivariate analysis consisted of Welch Two-Sample t-tests to compare the means of two groups and using the ‘aggregate’ function in R to examine counts of continuous variables for different categorical variables (5,29). I’ll start out with the aggregate function findings. The two main categories I was interested in for this was ‘cad’ and ‘sex’. This will allow me to divide up the data into ‘male’ and ‘female’, as well as ‘positive diagnosis’ and ‘negative diagnosis’, and examine different continuous variables for each category. The output I chose was a 6-number summary.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **rbp** |  | | cad | 0 | 1 | | Min | 94 | 100 | | 1st Qu. | 120 | 120 | | Median | 130 | 130 | | Mean | 129.175 | 134.635 | | 3rd Qu. | 140 | 145 | | Max | 180 | 200 | | |  |  |  | | --- | --- | --- | |  | **chol** |  | | cad | 0 | 1 | | Min | 126 | 131 | | 1st Qu. | 208.75 | 218 | | Median | 235 | 253 | | Mean | 243.4983 | 251.8540 | | 3rd Qu. | 268.25 | 284 | | Max | 564 | 409 | |
| |  |  |  | | --- | --- | --- | |  | **age** |  | | cad | 0 | 1 | | Min | 29 | 35 | | 1st Qu. | 44.75 | 53 | | Median | 52 | 58 | | Mean | 52.644 | 56.759 | | 3rd Qu. | 59 | 62 | | Max | 76 | 77 | | |  |  |  | | --- | --- | --- | |  | **cp (counts only)** |  | | cad | 0 | 1 | | cp.0 | 16 | 7 | | cp.1 | 40 | 9 | | cp.2 | 65 | 18 | | cp.3 | 39 | 103 | |

Since the CDC said chest pain (angina), age, blood pressure, and cholesterol may be contributing factors to CAD, those are the 4 variables I examined against ‘cad’. This table can be interpreted by looking at the row ‘cad’ which has values “0” and “1”. Recalling that 0 means negative diagnosis and 1 means positive diagnosis, each column can be explored. For example, the mean resting blood pressure (rbp) value for a person who has CAD is roughly 135. The other 3 variables can be interpreted similarly. I did a similar comparison with ‘sex’ but I used more variables than the 4 above.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **rbp** |  | | sex | Female | Male | | Min | 94 | 94 | | 1st Qu | 120 | 120 | | Median | 132 | 130 | | Mean | 133.3958 | 130.8806 | | 3rd Qu | 140 | 140 | | Max | 200 | 192 | | |  |  |  | | --- | --- | --- | |  | **chol** |  | | sex | Female | Male | | Min | 141 | 126 | | 1st Qu | 214.75 | 211 | | Median | 255 | 236 | | Mean | 262.2292 | 240.2438 | | 3rd Qu | 302.5 | 270 | | Max | 564 | 353 | |
| |  |  |  | | --- | --- | --- | |  | **mhr** |  | | sex | Female | Male | | Min | 96 | 71 | | 1st Qu | 142 | 132 | | Median | 157 | 150 | | Mean | 151.6042 | 148.6418 | | 3rd Qu | 165 | 166 | | Max | 192 | 202 | | |  |  |  | | --- | --- | --- | |  | **oldpeak** |  | | sex | Female | Male | | Min | 0 | 0 | | 1st Qu | 0 | 0 | | Median | 0.6 | 0.8 | | Mean | 0.8760 | 1.1413 | | 3rd Qu | 1.4 | 1.9 | | Max | 6.2 | 5.6 | |
| |  |  |  | | --- | --- | --- | |  | **age** |  | | sex | Female | Male | | Min | 34 | 29 | | 1st Qu | 49.75 | 47 | | Median | 57 | 55 | | Mean | 55.75 | 53.97 | | 3rd Qu | 63 | 60 | | Max | 76 | 77 | | |  |  |  | | --- | --- | --- | |  | **rbpcode** |  | | sex | Female | Male | | hyp | 33 | 65 | | norm | 29 | 68 | | phyp | 34 | 68 | |
| |  |  |  | | --- | --- | --- | |  | **normchol** |  | | sex | Female | Male | | normchol.0 | 14 | 35 | | normchol.1 | 82 | 166 | | |  |  |  | | --- | --- | --- | |  | **cad** |  | | sex | Female | Male | | cad.0 | 71 | 89 | | cad.1 | 25 | 112 | |

The last section of bivariate exploration was preforming t-tests for different variables. T-tests compare the means of two groups, and in R the ‘t.test’ function outputs a Welch Two-Sample t-test. The variables I explored at this step included sex&rbp, sex&chol, sex&mhr, cad&age, cad&rbp, cad&chol, cad&mhr, and cad&oldpeak. The null and alternative hypothesis for Welch’s t-test are as follows.

H0: The population means are equal

HA: The population means are not equal

Using an alpha value of 0.05, sex&chol, cad&age, cad&rbp, cad&mhr, and cad&oldpeak had p-values < 0.05. For these five combinations of variables examined, the null hypothesis can be rejected in favor of the alternative hypothesis. Interpreting sex&chol, it can be determined that the mean for cholesterol among ‘males’ and ‘females’ are not equal. Similar interpretations can be drawn from the other 4 combinations of variables. The remaining three combinations of variables, sex&rbp, sex&mhr, and cad&chol, didn’t have a p-value less than 0.05, therefore the null cannot be rejected for those tests.

**Outliers**

To determine outliers for each numerical variable, I created box plots as well as used the ‘boxplot.stats’ function in R to obtain index values or outliers (8).

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

The five numeric variables along with the index number containing the outlier is given below

|  |  |
| --- | --- |
| **Variable** | **Index #** |
| age | none |
| rbp | 10, 46, 77, 118, 168, 179, 210, 235, 243 |
| chol | 80, 84, 181, 189, 239 |
| mhr | 165 |
| oldpeak | 10, 187, 214, 240, 261 |

Since univariate outliers were found, the two approaches discussed in the results section were done twice; once with outliers and once with outliers removed.

Once exploratory analysis has been completed, I started working towards building a logistic regression model by stepwise processes. The first approach I took only utilized the binary variable ‘cad’ and numeric variables ‘age’, ‘rbp’, ‘chol’,’mhr’, and ‘oldpeak’. The foundation of this approach was based on a similar model by Penn State Eberly College of Science. (1).

A second approach I took utilized the ‘glm’ function in R which I discussed briefly in the introduction section. It also uses a stepwise process, but only attempts to model the patients CAD diagnosis being “1”, or positive (15). Since I previously checked associations as part of the exploratory analysis, I can include all variables in the first step of the model building procedures.

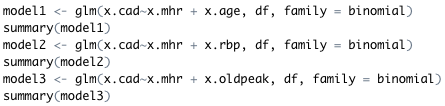
**Results**

*Approach One*

The first step in stepwise logistic regression is to set an alpha enter/exit value. There’s some debate over which value should be used for this, but I chose to use 0.15 per Penn State suggestion (1). Next, using the ‘glm’ function in R, I ran five different simple logistic regression models with each categorical variable against ‘cad’, these are the results.

|  |  |
| --- | --- |
| Model | P-value |
| cad~age | 0.000128 |
| cad~rbp | 0.00914 |
| cad~chol | 0.170 |
| cad~mhr | 2.28e-11 |
| cad~oldpeak | 3.25e-11 |

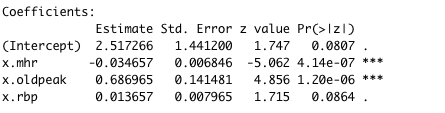
Looking at the p-values from each simple logistic regression model, ‘chol’ doesn’t meet the 0.15 threshold in order to be considered in future models, so it is omitted. Then, in order to work towards a multiple logistic regression model, the smallest p-value is chosen as the ‘best’ variable and has a permanent position in the equation. The second step consists of regressing ‘mhr’ and every other variable on ‘cad’ to create 3 different models, this is shown below.



The results from these 3 models is shown below.

|  |  |
| --- | --- |
| Variables | P-values |
| mhr  age | 1.95e-9  0.19511 |
| mhr  rbp | 3.08e-11  0.01422 |
| mhr  oldpeak | 4.96e-7  4.02e-7 |

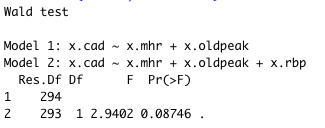
From these results, ‘age’ is omitted and ‘oldpeak’ is the best variable from this step. Since ‘age’ is omitted from the model, the only possible variable left to include is rbp. Therefore, the third and final step is to regress ‘cad’ against ‘mhr’, ‘oldpeak’, and then include ‘rbp’ and examine the p-value to decide whether or not it gets included in the final model. The results for this model are as follows.



Since ‘rbp’ p-value is less than 0.15, it can be kept in the model, but further testing on its significance will be done. Utilizing Wald’s test, I can compare the model that contains ‘rbp’ and the model that doesn’t. Wald’s test helps in examining whether a variables inclusion in a model is significant. The null and alternative hypothesis for Wald’s test are as follows.

H0: The variables of interest are equal to 0

HA: The variables of interest are not equal to 0



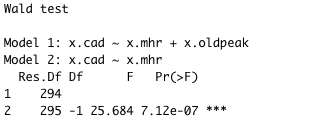
Since the p-value of the Wald test is greater than 0.05, the null hypothesis cannot be rejected. Therefore, adding the variable ‘rbp’ doesn’t significantly add anything to the model.

In order to further validate approach one, I then examined the AIC values for models of interest. AIC evaluates the model fit on the data it was calculated from and penalizes models for including too many predictor variables. A lower AIC value indicates a better fit (30). I started out by examining the AIC value from the full model, using all five numerical variables. A table of different models and respective AIC values are given below.

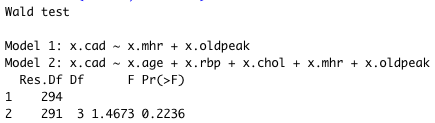
|  |  |  |
| --- | --- | --- |
| Model | Model | AIC Value |
| glm.age | cad~age | 398.25 |
| glm.rbp | cad~rbp | 406.88 |
| glm.chol | cad~chol | 412.03 |
| glm.mhr | cad~mhr | 355.97 |
| glm.oldpeak | cad~oldpeak | 354.48 |
| Model 1 | cad~mhr+age | 356.28 |
| Model 2 | cad~mhr+rbp | 351.75 |
| Model 3 | cad~mhr+oldpeak | 327.33 |
| Final Model | cad~mhr+oldpeak+rbp | 326.36 |
| Full Model | cad~age+rbp+chol+mhr+oldpeak | 328.85 |

Single variable models compared to two or three variables models, the AIC value generally decreases, which is a positive indicator of model improvement. Looking at ‘Model 3’, ‘Final Model’ and ‘Full Model’, there isn’t much difference between their AIC values. Even though ‘rbp’ met the p-value threshold to be included in the model during the stepwise process, it was determined not to significantly add to the model by the Wald test. More Wald tests can be used to compare the models where AIC value didn’t differ much.

Since it was determined that ‘rbp’ actually shouldn’t be in the model, I wanted to test whether or not Model 3 was any different from ‘glm.mhr’ (reference above table). Since ‘mhr’ was found in step one of the stepwise process for being the ‘best’ variable, I want to know whether or not the inclusion of ‘oldpeak’, from step two, significantly adds to the model.



The p-value for this test is less than 0.05, so it can be determined that including ‘oldpeak’ does significantly add to the model. The last thing I wanted to do was compare this two-variable model, with ‘mhr’ and ‘oldpeak’, and compare it to the full model with all five continuous variables.

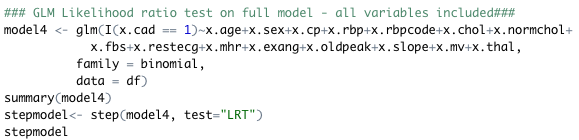


Because the p-value is not less than 0.05, it can be concluded that including ‘age’, ‘rbp’, and ‘chol’ does not add anything to the model. This was already partially determined as ‘age’ and ‘chol’ were removed in the stepwise process, but I wanted to further validate the findings. Utilizing the AIC values of different models as well as Wald tests, I can determine that the two variables that should be included in the final model are “mhr’ and ‘oldpeak’. Due to the presence of outliers in the dataset though, I need to run approach one again with outliers omitted and compare in order to truly determine a final model (14). This will be discussed later after approach two.

**Approach two**

The second approach also utilizes the ‘glm’ function in R, but instead it uses all the variables and only attempts to fit a model for CAD positive patients, coded ‘1’ in the dataset.

This was done with the following code in R.



Essentially, this preforms a stepwise search while utilizing likelihood ratio tests for different combinations of variables and arrives at a ‘best’ model. The model that this approach designed includes variables ‘sex’(male), ‘cp’ (levels 1,2,&3), ‘rbp’, ‘mhr’, ‘exang’(level 1),‘oldpeak’, ‘slope’(levels 1&2), ‘mv’(levels 1,2,&3), and ‘thal’(levels 1&2). Since much of this process is automated in R, a main indicator that the best model was arrived at is the AIC value are each step of the process. The beginning AIC value of the full model, containing all variables, was 229.12. Then at the final step of the process, the AIC value is 219.7, which is a good sign. Just like approach one, before drawing any final conclusions, a second process for each approach needs to be done where outliers are omitted from the dataset and models are compared.

**Outliers**

In order to assess whether or not outliers should be excluded from the model, I ran both approaches with the full data set, and then again with a dataset where the rows containing univariate outliers were omitted. The entire row was deleted if there existed an outlier for one of the numeric variables. This resulted in the deletion of 19 rows, which leaves 278 observations in the dataset. After completing the two approaches discussed above with the reduced dataset, I can conclude that there are some differences when outliers are excluded.

In approach one, ‘rbp’ remained below the alpha enter/exit value, 0.0864, in step three, which determined that it should be included in the model. Then using Wald tests, it was determined that there was no significant difference in whether ‘rbp’ was included in the model or not.

Using the dataset where outliers were deleted, the variable ‘rbp’ does not meet the requirements for model inclusion in step three of the stepwise process (p-value of 0.243), which is not less than 0.15. However, step one and two remained the same, ‘mhr’ was the best variable in step one and then ‘oldpeak’ was best in step two. Using the dataset where outliers are omitted results in a logistic regression model with only ‘mhr’ and ‘oldpeak’. The same determination was made with the dataset including outliers by implementing a Wald test after the stepwise process. Due to this, I would conclude using the dataset where outliers were omitted results in the best model for approach one.

Moving onto the second approach, I used the same code as shown above, and instead used the dataset of 278 observations where outliers were deleted. There were two main differences between the model with outliers and without. When using the dataset with outliers removed, ‘exang’(level 1) was no longer in the final model, and ‘chol’ & ‘normchol’(level 1) were included. Due to research on the subject of heart disease and knowing the key role that cholesterol plays in the development of CAD, I would also conclude that omitting outliers from the dataset results in an overall better model for approach two.

To further conclude and validate variable selection, I used the anova() function in R, with the test set to “Chisq”, on the models derived from both approaches, as well as the models derived from the dataset with outliers removed. The results are shown below

**Approach 1**

With Outliers Outliers Removed

|  |  |
| --- | --- |
|  |  |

**Approach 2**

With Outliers Outliers Removed

|  |  |
| --- | --- |
|  |  |

This ANOVA uses the chi-square distribution and measures the deviance for each parameter. The higher the deviance would indicate a worse fit. The p-values given tell me which predictors have a significant effect on the probability of have ‘cad’ being “1”. The results in the first approach make sense and indicate that only variables ‘mhr’ and ’oldpeak’ should be included in the logistic regression model when examining only numeric data against ‘cad’. Examining the ANOVA for approach two with outliers, all the variables have a p-value less than 0.05. However, on the outliers removed side, ‘slope’ and ‘normchol’ do not have a p-value of less than 0.05. These values can be interpreted as such: When controlling for all other variables, ‘sex’ does significantly predict a person having a response of ‘1’ for variable ‘cad’. All variables with p-values less than 0.05 can be interpreted the same way. These conclusions cannot be made for the variables whose p-value is above 0.05 (2).

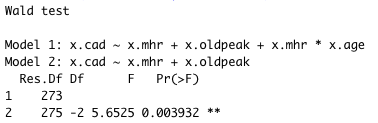
Knowing cholesterols significance in development of CAD, I would opt to use the dataset that does not include outliers, since ‘chol’ became significant when doing so. After comparing the results of both approaches, I would determine that using the dataset with outliers removed is best. From now on, I will only continue with the outliers removed dataset.

Since two of the variables on the outliers removed section of the ANOVA chi-square test resulted in p-values greater than 0.05, Wald tests can be used yet again to examine those specific variables. Specifically, I want to compare the model from approach two with ‘normchol’ and ‘slope’. I ran a Wald test for the model derived from approach two (outliers removed) with each variable that was included, and it was determined that ‘normchol’, ‘slope’, and ‘mhr’ do not significantly add to the model, therefore they can be excluded from the final model.

One last thing that needs to be checked before deriving a final model is possible interaction effects.

**Interactions**

While a single variable might not have a significant impact on the model, an interaction between two variables possibly could. Specifically, in approach one, I noticed that many variables that are deemed risk factors by the CDC were not included in the model. Overall, I tested nine different interactions for approach one, but the only interaction the was significant was that between ‘age’ and ‘mhr’. This interaction makes sense, as maximum heart rate is calculated based on a person’s age (28). When comparing the model with the interaction term age\*mhr and the model without the interaction, the AIC value decreases from 305.29 to 296.34, which indicates a better fit with the inclusion of the interaction term. To further test the significance of adding the interaction term, I can use the Wald test.



Interactions should also be examined for approach two, but due to page limitations they were not discussed.

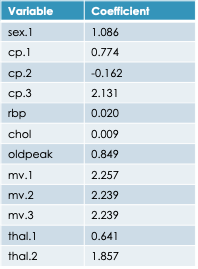
Therefore, the final selection of variables for the multiple logistic model for approach one without outliers, results in using ‘mhr’,‘oldpeak’, and the interaction effect between ‘age’ and ‘mhr’, which by default also includes ‘age’ as part of the final model. The final variable selection for approach two results in the use of variables ‘sex’(1), ‘cp’(1,2,3), ‘rbp’, ‘chol’, ‘oldpeak’, ‘mv’(1,2,3), and ‘thal’(1,2).

The equations for approach one is shown below, but approach took up too much space, but it would use a similar equation, just Bnxn from 0 through 12 and no interaction effect.

These variables can be interpreted as such. For every one-unit change in ‘mhr’, the log-odds of ‘cad’ increases by a factor 0.1928. Or I could transform the values into the odds ratio and interpret it differently. To convert the value to odds ratio, I simple did eB1 which results in 1.2126. This means that for every one unit increase in ‘mhr’, the odds of having ‘cad’ be positive increases by a factor of 1.2126. The other variables in the equation can be interpreted similarly.

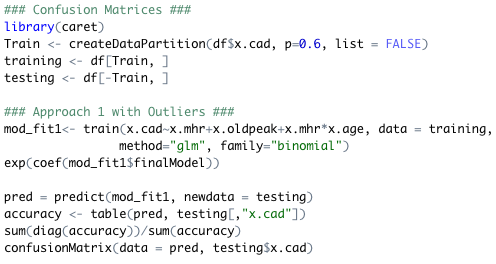
The coefficient values for approach two are given in the following table, the intercept value is

-9.3844.



An interpretation of one of the categorical variables would be something like this. When you’re male (sex.1), compared to female (sex.0), the log-odds of cad being ‘1’ changes by a factor of 1.086. Using the odds-ration, e1.086 = 2.96, which means that the odds of ‘cad’ being ‘1’ (positive) when you are male, is 2.96 times that of the odds if you’re female.

To get a better understanding of how well this model predicts the probability of someone having CAD, a confusion matrix was calculated (7). This compares the observed Y values to the predicted Y values and will give a percentage of how accurately it predicts correctly. The confusion matrix for both the model with outliers included and outliers excluded are given. Note, these confusion matrices were calculated with a testing and training set of the data, so it varies each time it’s ran. The code for implementing this is shown below for approach one with outliers.



This was done for both approaches, with and without outliers, and it summarized below. 0 and 1 in the matrix below don’t indicate CAD diagnosis, 0 simply means “True” and 1 means “False”. I wanted to show the accuracy differences between approaches when outliers were included versus omitted. A general layout of a confusion matrix is provided below.

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | True Positives | False Negatives |
| Negative | False Positives | True Negatives |

Accuracy is calculated along the diagonal, ‘True Positives’ + ‘True Negatives’ / All. Since this confusion matrix was calculated with a testing and training set, it varies slightly every time. One of the runs is provided below.

**Approach 1**

***With Outliers*  *Without Outliers***

*Reference*

*Reference*

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 54 | 16 |
| 1 | 10 | 38 |

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 49 | 16 |
| 1 | 15 | 38 |

*Prediction*

*Prediction*

* Accuracy: 0.7373
* 95% CI: 0.6483, 0.814)
* No Information Rate: 0.5424
* P-Value: 1.042e-5
* Sensitivity: 0.7656
* Specificity: 0.7037
* Accuracy: 0.7808
* 95% CI: (0.6686, 0.8692)
* No Information Rate: 0.5479
* P-Value: 3.035e-5
* Sensitivity: 0.8250
* Specificity: 0.7273

**Approach 2**

***With Outliers*  *Without Outliers***

*Reference*

*Reference*

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 54 | 16 |
| 1 | 10 | 38 |

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 51 | 12 |
| 1 | 13 | 42 |

*Prediction*

*Prediction*

* Accuracy: 0.7881
* 95% CI: (0.7033, 0.858)
* No Information Rate: 0.5424
* P-Value: 2.396e-8
* Sensitivity: 0.7969
* Specificity: 0.7778
* Accuracy: 0.9178
* 95% CI: (0.8296, 0.9692)
* No Information Rate: 0.5479
* P-Value: 7.396e-8
* Sensitivity: 0.9250
* Specificity: 0.9091

Plots to showcase the accuracy of the above models (only without outliers) are also given below.

**Approach 1 Approach 2**

|  |  |
| --- | --- |
|  |  |

Referring to the confusion matrix provided for approach one and two without outliers, it says that the model form approach one accurately predicts CAD diagnosis 78.08% of the time. For approach two, 91.78% of the time. Examining the 95% confidence intervals, approach one is accurate 66%-86% of the time, while approach two is accurate 82-96% of the time.

Comparing these results to the visuals provided above, the percentages look as though they match. Where teal represents having CAD = ‘1’, and orange represents not having CAD = ‘0’, it’s a good indicator that a majority of teal is closer to 1 and majority of orange is closer to 0.

**Conclusion**

Although stepwise regression doesn’t necessarily guarantee the best model, the two models derived in this study do a sufficient job at predicting CAD diagnosis. There are potential other models that are equally good. As previously stated, interaction effects should be examined for approach two as well.

Due to the accuracy percentages presented in the confusion matrix, I would conclude that the model derived in approach two provides a better fit.

With more time and resources, interaction effects should be examined more thoroughly for both approaches, especially approach two. Consideration of utilizing more complex machine learning techniques would also be appropriate. Examining a larger dataset would also be useful.

**References**

1. “10.2 - Stepwise Regression: Stat 501.” *PennState: Statistics Online Courses*, https://online.stat.psu.edu/stat501/lesson/10/10.2.
2. *23. BINOMIAL ANOVA - University of Alberta*. https://sites.ualberta.ca/~lkgray/uploads/7/3/6/2/7362679/23\_-\_binomial\_anova.pdf.
3. “8. Ventricular Hypertrophy.” *ECG Learning Center - An Introduction to Clinical Electrocardiography*, https://ecg.utah.edu/lesson/8.
4. “Achievements in Public Health, 1900-1999: Decline in Deaths from Heart Disease and Stroke -- United States, 1900-1999.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, https://www.cdc.gov/mmwr/preview/mmwrhtml/mm4830a1.htm.
5. “Aggregate: Compute Summary Statistics of Data Subsets.” *RDocumentation*, https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aggregate.
6. “Angina (Chest Pain).” *Www.heart.org*, https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain.
7. Atmathew. “Evaluating Logistic Regression Models: R-Bloggers.” *R*, 18 Aug. 2015, https://www.r-bloggers.com/2015/08/evaluating-logistic-regression-models/.
8. “Boxplot.stats: Box Plot Statistics.” *RDocumentation*, https://www.rdocumentation.org/packages/grDevices/versions/3.6.2/topics/boxplot.stats.
9. Cherngs. “Heart Disease Cleveland UCI.” *Kaggle*, 29 Mar. 2020, https://www.kaggle.com/cherngs/heart-disease-cleveland-uci?select=heart\_cleveland\_upload.csv.
10. “Cholesterol Levels: What You Need to Know.” *MedlinePlus*, U.S. National Library of Medicine, 2 Oct. 2020, https://medlineplus.gov/cholesterollevelswhatyouneedtoknow.html.
11. “Coronary Artery Disease.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 19 July 2021, https://www.cdc.gov/heartdisease/coronary\_ad.htm.
12. Finkelhor, Robert S., et al. “The St Segment/Heart Rate Slope as a Predictor of Coronary Artery Disease: Comparison with Quantitative Thallium Imaging and Conventional ST Segment Criteria.” *American Heart Journal*, vol. 112, no. 2, 1986, pp. 296–304., https://doi.org/10.1016/0002-8703(86)90265-6.
13. “Fluoroscopy: Medlineplus Medical Test.” *MedlinePlus*, U.S. National Library of Medicine, 12 Nov. 2021, https://medlineplus.gov/lab-tests/fluoroscopy/.
14. Frost, Jim, et al. “Guidelines for Removing and Handling Outliers in Data.” *Statistics By Jim*, 5 Apr. 2021, https://statisticsbyjim.com/basics/remove-outliers/.
15. “Generalized Linear Models in R.” *Social Science Computing Cooperative*, 1 May 2021, https://sscc.wisc.edu/sscc/pubs/glm-r/.
16. “GLM: Fitting Generalized Linear Models.” *RDocumentation*, https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm.
17. Hajar, Rachel. “Risk Factors for Coronary Artery Disease: Historical Perspectives.” *Heart Views*, vol. 18, no. 3, 2017, p. 109., https://doi.org/10.4103/heartviews.heartviews\_106\_17.
18. “Heart Disease Facts.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 27 Sept. 2021, https://www.cdc.gov/heartdisease/facts.htm.
19. “High Blood Pressure Symptoms and Causes.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 18 May 2021, https://www.cdc.gov/bloodpressure/about.htm.
20. Kashou, Anthony H. “St Segment.” *StatPearls [Internet].*, U.S. National Library of Medicine, 11 Aug. 2021, https://www.ncbi.nlm.nih.gov/books/NBK459364/.
21. Kassambara, and Michael U. “Logistic Regression Assumptions and Diagnostics in R.” *STHDA*, 11 Mar. 2018, http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/.
22. Kutner, Michael H., et al. “Chapter 9 and 14.” *Applied Linear Regression Models*, Mac-Graw Hill/Irwin, New York, 2004.
23. “Multiple Logistic Regression Analysis.” *Multiple Logistic Regression Analysis*, https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704\_multivariable/bs704\_multivariable8.html.
24. Nelson, Jennifer. “Thallium Stress Test: Purpose, Procedure, and Risks.” *Healthline*, Healthline Media, 17 Sept. 2018, https://www.healthline.com/health/thallium-stress-test.
25. “PH717 Module 9 - Correlation and Regression.” *The Correlation Coefficient (r)*, https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html.
26. “Provisional Mortality Data - United States, 2020.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 17 June 2021, https://www.cdc.gov/mmwr/volumes/70/wr/mm7014e1.htm.
27. “Selecting between Parametric and Non-Parametric Analyses.” *Statistics Solutions*, 23 June 2021, https://www.statisticssolutions.com/selecting-between-parametric-and-non-parametric-analyses/.
28. “Target Heart Rate and Estimated Maximum Heart Rate.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 14 Oct. 2020, https://www.cdc.gov/physicalactivity/basics/measuring/heartrate.htm.
29. “Using T-Tests in R.” *UCB Statistics*, https://statistics.berkeley.edu/computing/r-t-tests.
30. Zach. “What Is Considered a Good AIC Value?” *Statology*, 20 May 2021, https://www.statology.org/what-is-a-good-aic-value/.