# 621-FinalProject

Ahmed Elsaeyed

2022-11-30

# Probability of Cardiovascular Disease

This research centers around building a model that predicts the probability of cardiovascular disease given some prior health information.

# Abstract

For our final assignment, we will be looking at a dataset of physical characteristics, behavioral characteristics, and health results of patients aged 30-59 in the 1940's in the United States. The data was gathered by the Framingham Heart Study, which is a long-term ongoing study. 5209 subjects were monitored to see if cardiac health can be influenced by lifestyle and environmental factors. It is mainly due to studies like this one, and indeed this one in particular due to its large scope, that we take that conclusion for granted. Before this study, the contributing factors of cardiac health were not clear at all.

A result of this study is the 10-year cardiovascular risk score was developed, which is an estimation of how likely it is that someone will have a coronary disease based on the findings of the study. The objective of this project is to build a model to predict this to take their research a step further.

# Keywords - Coffy

# Introduction - Coffy

## Literature Review

Framingham Study is a study initiated by the United State Public Health Service, that started in 1948 and has been ongoing since. The origin of the study has been closely linked with the cardiovascular health of President Franklin D. Roosevelt and his death from hypertensive heart disease and stroke in 1945. It aims to investigate the epidemiology and factors that are involved in the development of cardiovascular disease. The plan of the study was to track a large cohort of patients over time. The study initially recruited the original cohort of 5209 participants from the town if Framingham, MA. The age group for this study ranges from 30-59 years old. The patients were given questionnaires and exams every 2 years and these questions and exams expanded with time. This study was continued over three generations of the original participants.

The data collected by the study was used to develop the Framingham Risk Score. The Framingham Risk Score is a sex-specific algorithm that is used to estimate the 10 year cardiovascular risk of an individual. This score gives an estimate

# Methodology

Our methodology for creating a predictive model for future coronary heart disease begins with cleaning and preprocessing our subject dataset. The dataset itself contains 16 fields, one of which is the target: `TenYearCHD`. While the dataset was relatively complete to begin with, we needed to employ multiple operations in order to transform it into a format which subsequent models could interpet. In this vain, we addressed missing values, multicollinearity, outliers, and mis-shaped data using a combination of self-developed functions and transformations. After cleaning / preprocessing, we split the new dataframe into training and testing sets in order to create and test various machine learning models. While developing models, we decided to try multiple approaches, including Binary Logistic Regression, standard and with log transforms, and Step-AIC. After each model, we calculated the relvant performance metrics and stored them in a tracker. Ultimately, our selection criteria was heavily influenced by the models' recall and f1 scores. With medical diagnosis, it's imperative that individuals with an illness be properly classified.

# Experimentation and Results

```
[1] 4240   16
```

# Data Exploration

Our data set had 4,240 observations and 16 variables. `Sex`, `education`, `currentSmoker`, `BPMeds`, `prevalentStroke`, `prevalentHyp`, `diabetes` and `TenYearCHD` in the analysis carried out in this report are categorical variables, while the rest are numeric. The histograms for the numeric variables revealed that `BMI`, `glucose`, and `totChol` were skewed, and a Box-Cox transformation was undertaken for these variables in order to create new variables that contained the transformed versions of `BMI`, `glucose`, and `totChol`. An analysis into the correlation was done in order to remove variables that were highly correlated with another. Also, the data had several observations with missing data. The methodology that was undertaken to account for these missing values is shown in the "*Missing Values*" section. For comparison purposes, two different datasets were created; one containing the original data with observations with missing values omitted, and another containing manipulated data, where the data went through several transformations as explained in this report.

```
    Sex              age           education    currentSmoker
```

# Multicollinearity Analysis

Correlation refers to the strength of a relationship between two or more variables. Problems arise when there significant correlation between two or more predictor variables within a dataset, as this can lead to solutions that are wildly varying and possibly numerically unstable and there is redundancy between predictor variables. One solution to this is to drop one variable that is highly correlated with the other. Calkins explains that we can describe the correlation values between variables in the follwing manner:

- Correlation coefficients between 0.9 and 1 = very highly correlated
- Correlation coefficients between 0.7 and 0.9 = highly correlated
- Correlation coefficients between 0.5 and 0.7 = moderately correlated
- Correlation coefficients under 0.5 = low correlation

The correlations between the variables in the dataset were calculated. What was found was that the vast majority of the variables had a correlation that was less than 0.25. `sysBP` and

# Missing Values

The dataset itself contains several observations containing missing values. The percentage of missing data for each of the predictor variables is shown in Table 1.

| Variable Name | Percentage of Missing Data |
|---|---:|
| education | 2.48 |
| cigsPerDay | 0.68 |
| BPMeds | 1.25 |
| totChol | 1.18 |
| BMI | 0.45 |
| heartRate | 0.02 |
| glucose | 9.15 |

*Table 2: Percentage of missing data for each of the variables that had missing data.*

Ultimately, it was decided to remove the observations with the missing values instead of imputing. The reason for this is explained

# Removing Outlier Values

The boxplots, summaries, and histograms that were generated revealed that some of these heart rates, diastolic, and systolic values as well as total cholesterol do not seem realistic. Imputing the values would not be appropriate because this is medical data and would harm the authenticity of the dataset, and so we have decided to simply remove the bad data.

We conducted research on suspected variables to further examine the authenticity of the data points in the dataset. Using multiple credited medical sources we created plausible ranges for each variable and omitted extreme or impossible values. Brief discussions and analysis of the process for each variable are below.

**Heart Rate**

An article by Cardiologist Jim Liu of Ohio State University states a normal resting heart rate, measured when the person is not exercising, is between 60 and 100 bpm (beats per minute). The further you move from 100 bpm increases the level of risk on the high sided. On the decreasing side going lower than 60 deviates

# Box-Cox Transformation for Skewed Variables

For the variables that exhibited skewness as explained in the "*Data Exploration*" section of this paper, A Modern Approach to Regression with R explains the following:

> *. . . if the skewed predictor can be transformed to have a normal distribution conditional on Y, then just the transformed version of X should be included in the logistic regression model. (Sheather, 284)*

In order to address this skewness and attempt to normalize these features for future modeling, we employed Box-Cox transformations. Because some of these values include 0, we replaced any zero values with infinitesimally small, non-zero values.

The $\lambda$'s that were used to transform the skewed variables are shown on Table 3.

| Column Name | $\lambda$ |
|---|---:|
| BMI | -0.309 |
| glucose | -1.279 |

# Split Data Into Testing and Training

The data was into testing and training subsets such that 70% of it will be used to train, and 30% to test. The first row shows the split for the testing data while the second row shows the split for the training data. Tables 4 and 5 show the distributions of the testing and training data using the original and modified datasets.

| | Number of Observations where $\texttt{TenYearCHD} = 1$ | Number of Observations where $\texttt{TenYearCHD} = 0$ |
|---|---|---|
| Test | 930 | 167 |
| Train | 2171 | 390 |

*Table 4: Distribution of testing and training data using original dataset*

| | Number of Observations where $\texttt{TenYearCHD} = 1$ | Number of Observations where $\texttt{TenYearCHD} = 0$ |
|---|---|---|

# Building Models

# Binary Logistic Regression Model with Original Data

A binary logistic regression model was generated using the original dataset. The observations with missing values were omitted. The final binary logistic regression takes on the following form:

$$
\begin{gathered}
logit(p) = -9.155 + (0.504 * Sex_{male}) + (0.066 * \\
(-0.290 * education_2) + (-0.223 * education_3) + (-0.120 * educati \\
(0.086 * currentSmoker_{Yes}) + (0.019 * cigsPerDay) + (0.123 * BPMe \\
(0.982 * prevalentStroke_1) + (0.091 * prevalentHyp_1) + (-0.435 * diabete \\
(0.002 * totChol) + (0.018 * sysBP) + (-0.005 * diaBP) + (0.02 * B \\
(-0.005 * heartRate) + (0.01 * gl
\end{gathered}
$$

*Equation 1: Binary logistic regression model using original data.*

In this model, the variables that had a p-value greater than 0.05 were `Sex` `age`, `cigsPerDay`, `sysBP` and `glucose`. The sex of the person definitely seems to play a significant factor in determining whether or not a person will develop heart disease. Figure XX revealed that men that were surveyed on average smoked more

# Binary Logistic Regression Model with Modified Data

A binary logistic regression model was generated using the modified dataset. The observations with missing values were omitted. The final binary logistic regression model that used modified data takes on the following form:

$$
\begin{aligned}
logit(p) = -128.3 + (0.52 * Sex_{male}) + (0 \\
(-0.105 * education_2) + (-0.55 * education_3) + (-0.074 * \\
(0.277 * currentSmoker_{Yes}) + (0.013 * cigsPerDay) + (0.091 * \\
(0.651 * prevalentStroke_1) + (0.145 * prevalentHyp_1) + (-0.026 * \\
(0.009 * totChol) + (0.013 * sysBP) + (0.18 * BMI) + (-0.009 * \\
(0.013 * glucose) + (-13 * tf\_BMI) + (-140.9 * tf\_glucose) + (-1.39 *
\end{aligned}
$$

*Equation 2: Binary logistic regression model using modified data.*

The parameters in this model that had p-values less than 0.05 were `Sex`, `age`, `education_3`, `sysBP` and `glucose`. None of the Box-Cox transformed variables had p-values less than 0.05. We see the same variables (except `cigsPerDay`) that were statistically

# Step-AIC Binary Linear Regression Model with Original Data

Step AIC works by deselecting features that negatively affect the AIC, which is a criterion similar to the R-squared. It selects the model with not only the best AIC score but also a model with less predictors than the full model, since the full model may have predictors that do not contribute or negatively contribute to the model's performance. The direction for the Step AIC algorithm was set to `both`, because this implements both forward and backward elimination in order to decide if a predictor negatively affects the model's performance. The final step-AIC binary logistic regression model that used the original data takes on the following form:

$$logit(p) = -9.589 + (0.503 * Sex_{male}) + (0.072 * age) + (0.02 * cigsPe$$

$$(1.004 * prevalentStroke_1) + (0.017 * sysBP) + (0.023 * BMI) + (0.009 *$$

*Equation 3: Step-AIC binary logistic regression model using the original data.*

The p-values for this model, which is far more parsimonious than

## Step-AIC Binary Logistic Regression Model with Modified Data

The final step-AIC binary logistic regression model that used the modified data takes on the following form:

$$\begin{aligned} logit(p) = -14.41 + (0.52 * Sex_{male}) + (0.073 * age) + (-0.09 * educatio \\ (-0.056 * education_4) + (0.02 * cigsPerDay) + (0.0 \\ (0.009 * gl \end{aligned}$$

*Equation 4: Step-AIC binary logistic regression model using the modified data.*

The p-values for this model, which is far more parsimonious than the original binary logistic regression model that used all of the predictors, were all below 0.05 except for `Intercept`, `education2`, `education4`, `BMI` and `tf_BMI`; 3 more than the amount of predictors that had values above 0.05 for the step-AIC binary logistic regression model using the original predictors.

# Model Selection

The various metrics that were used to determine the best model are shown in Table 6. Figure XX is a bar chart which plots all of the metrics for all of the models. Table 6 and Figure xx reveal that the accuracy, precision, recall, AUC, and F-score are higher when the original data was used as opposed to the modified. Table 6 revealed that the step-AIC binary logistic regression model using the original data has the second highest AIC, but it also has the highest precision and F-score out of all of the models. Taking all of this into consideration, and the fact that the step-AIC binary logistic regression model using the original data is the most parsimonious out of all of the models, this model is the best performing model out of all of them. The AUC curve for the step-AIC binary logistic regression with original data is provided in Figure xx.

| Model | Precision | Recall | AIC |
|---|---|---|---|
| Bin. Log. w/ Original Data | 0.68 | 0.11 | 1939.25 |
| Bin. Log. w/ Modified Data | 0.67 | 0.02 | 1507.5 |
| Step AIC Bin. Log. w/ Original Data | 0.7 | 0.11 | 1928.55 |
| Step AIC Bin. Log. w/ Modified Data | 0.67 | 0.02 | 1496.61 |

# Discussion and Conclusions

In this paper, 4 different binary logistic regression models were generated in order to predict the 10-year risk of chronic heart disease. The results from the analysis carried out in this report indicate that the transformation of the skewed variables to a normal distribution and the removal of outliers resulted in worse model performance when comparing the metrics to these models and the models that used the original unaltered dataset. With that being said, the AUC's for all of the models were relatively the same. The final model that was selected in order to predict the 10-year risk of chronic heart disease was the best performing and the most parsimonious. We were able to test the validity of this particular model from when the data was split into testing and training datasets.

# References

Center for Drug Evaluation and Research. (2021a, January 21). High Blood Pressure– Understanding the Silent Killer. U.S. Food And Drug Administration. https://www.fda.gov/drugs/special-features/high-blood-pressure-understanding-silent- killer

Center for Drug Evaluation and Research. (2021b, January 21). High Blood Pressure– Understanding the Silent Killer. U.S. Food And Drug Administration. https://www.fda.gov/drugs/special-features/high-blood-pressure-understanding-silent- killer

Framingham Study | Boston Medical Center. (n.d.). https://www.bmc.org/stroke-and-cerebrovascular-center/research/framingham-study

High cholesterol - Symptoms and causes. (2021, July 20). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/symptoms- causes/syc-20350800

Liu, J., MD. (2022a, July 19). What's a dangerous heart rate? What's a Dangerous Heart Rate? | Ohio State Health & Discovery.

# Appendix

FIGURES AND CODE GO HERE.