Logistic Regression on South African Heart Dataset: Predicting Coronary Heart Disease Through Risk Factors

Theodore Fitch
Department of Data Analytics, University of Maryland Global Campus
DATA 630: Machine Learning
Dr. Ami Gates
June 22nd, Summer 2021

Introduction:

The purpose of this analysis is to use known factors about a person's medical health to predict whether they have coronary heart disease (CHD) using a logistic regression model. Heart disease is the leading killer in the United States with one person dying every 36 seconds (C.D.C., n.d.). Well over 350,000 people died in one year alone and 20% of all CHD deaths occur in individuals less than 65 years of age (Mozaffarian, 2015). Most people have been personally affected by CHD whether a family member or a friend. Unlike many genetic diseases, CHD is preventable and can be curbed with the right interventions and life choices. These facts alone are motivating to understand the risk factors and make lifestyle changes in order to increase longevity.

Heart disease in general is endemic, and CHD is one of the most common forms of heart related disease. This is because it is not genetic or randomly occurring but gradually occurs to most people over time (to different degrees based on the risk factors). Coronary arteries supply oxygenated blood to the heart ensuring the heart is able to continuously function. However, the buildup of fat (plaque) in the coronary arteries occurs over one's lifetime which restricts blood flow to the heart. This eventually leads to a significant lack of oxygenated blood (ischemia) to the heart. In severe cases, heart attack (myocardial infarction) will occur which is when an artery becomes blocked (Torpy et al., 2009). Tissue not receiving oxygen will die quickly. If the heart cannot receive oxygen, it is a medical emergency, and a person will die without medical treatment.

CHD is one of the most well-studied diseases. The risk factors are well characterized: being over 40 for men; being over 45 for women; male sex; family history of CHD; smoking; hypertension (high systolic blood pressure); diabetes; obesity; non-healthy cholesterol levels

(high triglycerides, high LDL, low HDL, high total); being inactive; and accumulation of abdominal fat (Torpy et al., 2009). Interestingly, there is an observed protective effect observed with the consumption of alcohol so long as a person does not excessively drink (Marmot, 2001). Consequently, the well-known prevention activities one can proactively take to prevent CHD are: refusing to smoke; treating high blood pressure; controlling blood sugar for those who have diabetes; loss excess fat; maintain a healthy BMI (body mass index); sustain a healthy diet including eating plenty of high fiber fruits and vegetables, whole grains, and lean meats; and limiting intake of cholesterol, animal lipids, and sugar (Torpy et al., 2009).

Thus, the purpose of this analysis is to take a CHD dataset and determine risk factors within that dataset using a logistic regression model. Whereas linear regression looks to predict exact values of a variable, the logistic regression method is excellent at predicting which category a datapoint belongs to based on other variables. The observed risk factors will be compared against the literature determined risk factors as one way to assess the accuracy of the models.

Analysis and Model Demonstration:

Data Information:

The dataset used retrieved from the KEEL website which is a subset of Soft Computing Intelligent Information Systems which is a subset of University of Granada (KEEL, n.d.). This is a real-world dataset (not generated). It will henceforth be referred to as “SAHeart”.

Exploratory Data Analysis:

```
> setwd("C:/Users/soari/Documents/Assignments/Data Analytics/UMGC/Summer 2021 Data 630/Assignment
 2")
> w=read.table("SAheart.csv",sep=",",header=TRUE, as.is = FALSE)
> str(w)
'data.frame':   462 obs. of  10 variables:
 $ sbp       : int  160 144 118 170 134 132 142 114 114 132 ...
 $ tobacco   : num  12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
 $ ldl       : num  5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
 $ adiposity: num  23.1 28.6 32.3 38 27.8 ...
 $ famhist   : Factor w/ 2 levels "Absent","Present": 2 1 2 2 2 2 1 2 2 2 ...
 $ typea     : int  49 55 52 51 60 62 59 62 49 69 ...
 $ obesity   : num  25.3 28.9 29.1 32 26 ...
 $ alcohol   : num  97.2 2.06 3.81 24.26 57.34 ...
 $ age       : int  52 63 46 58 49 45 38 58 29 53 ...
 $ chd       : int  1 1 0 1 1 0 0 1 0 1 ...
```

Figure 1. Structure of the SAHeart dataset shows there are 10 variables and 462 rows of data.

SAHeart contains 462 rows of data with 10 variables (Figure 1). CHD is the factor of interest to predict, and it is listed as a binary variable (1 = the person has CHD, and 0 = the person does not have CHD). This was immediately transformed from a numeric variable type to a factor variable type. There are 3 integer variables: SBP (Systolic Blood Pressure), Type A (a measure of personality and lifestyle), and age. Type A Behavior Pattern (TABP) is described as behavior characterized by the following: competitiveness, ambition, work-drive, time-consciousness, and aggression (Petticrew et al., 2012). This was originally described as a predictor of risk for CHD in the 1950's and studies corroborated this through the 1980's; however, more recent studies did not show that TABP is not a concrete predictor of CHD (Schwalbe, 1990; Rosenman & Chesney, 1980; Baker & Krantz, 2007). It is now ambiguous whether TABP is a consistent causal factor of CHD and why it was found to be so in some studies but not in others. There are 5 numeric variables: Tobacco (cumulative tobacco consumed in KG), LDL (low density lipoprotein cholesterol), Adiposity (measure of level of adipose/fat tissue), Obesity (measured in body mass index [BMI]), and Alcohol. How alcohol is measured is not listed (KEEL, n.d.) however it ranges from 0 to 147. Therefore, it is likely not a measure of whole drinks over time. It could be an average number over a year (or some other measure that would result in a decimal). All variables were explored to understand their distributions (Figure

2). The two factor variables, CHD and family history, showed negative values (CHD absent and CHD absent from family history) in a ratio of 2:1 and 1.4:1 respectively. The alcohol and tobacco variables showed means significantly greater than their medians indicating they had right skewed distributions. Each were right skewed and contained 3 outliers (Figure 5, 6).

```
> summary(w)
      sbp             tobacco            ldl              adiposity       famhist
 Min.   :101.0   Min.   : 0.0000   Min.   : 0.980   Min.   : 6.74   Absent :270
 1st Qu.:124.0   1st Qu.: 0.0525   1st Qu.: 3.283   1st Qu.:19.77   Present:192
 Median :134.0   Median : 2.0000   Median : 4.340   Median :26.11
 Mean   :138.3   Mean   : 3.6356   Mean   : 4.740   Mean   :25.41
 3rd Qu.:148.0   3rd Qu.: 5.5000   3rd Qu.: 5.790   3rd Qu.:31.23
 Max.   :218.0   Max.   :31.2000   Max.   :15.330   Max.   :42.49
     typea          obesity          alcohol            age          chd
 Min.   :13.0   Min.   :14.70   Min.   :  0.00   Min.   :15.00   0:302
 1st Qu.:47.0   1st Qu.:22.98   1st Qu.:  0.51   1st Qu.:31.00   1:160
 Median :53.0   Median :25.80   Median :  7.51   Median :45.00
 Mean   :53.1   Mean   :26.04   Mean   : 17.04   Mean   :42.82
 3rd Qu.:60.0   3rd Qu.:28.50   3rd Qu.: 23.89   3rd Qu.:55.00
 Max.   :78.0   Max.   :46.58   Max.   :147.19   Max.   :64.00
```

Figure 2. Summary statistics of each variable show tobacco and alcohol have means significantly larger than their medians.

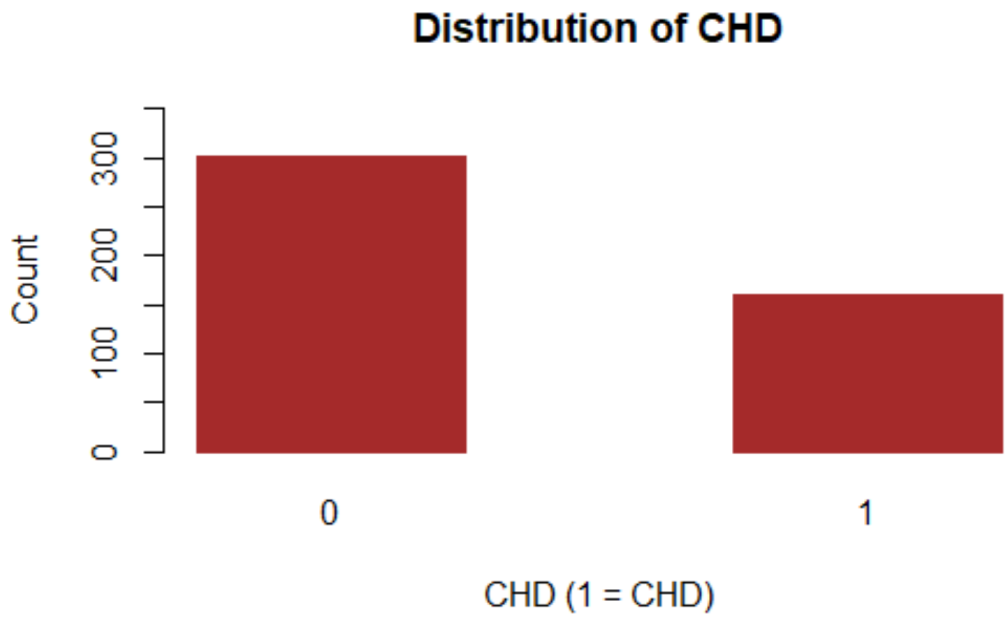

Figure 3. Distribution of CHD presence shows there are 160 CHD positive individuals and 302 negative individuals.

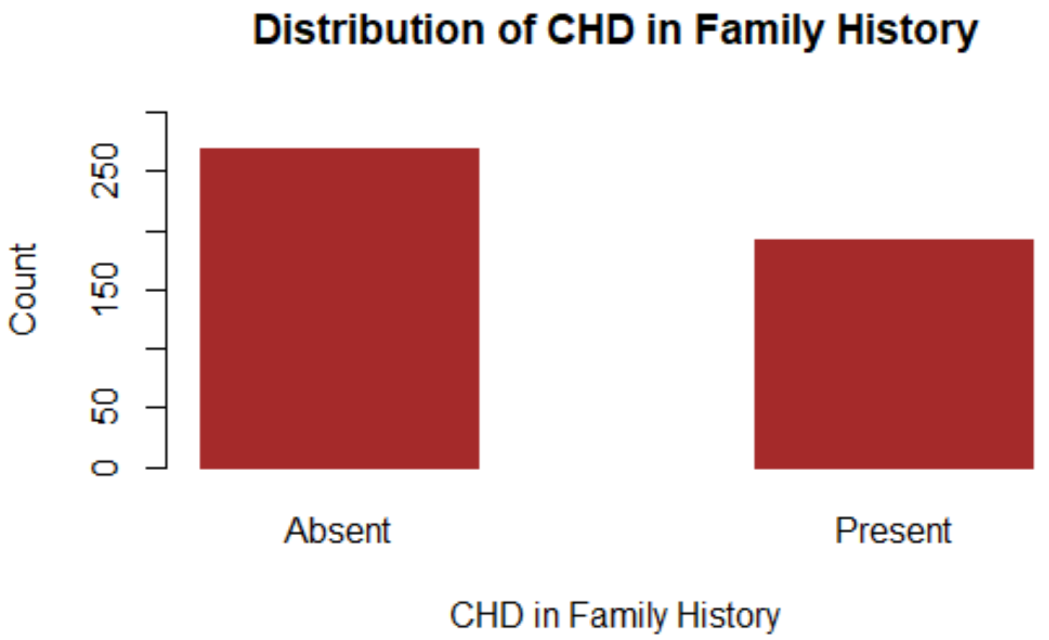

Figure 4. Distribution of CHD in family history shows there are 270 individuals with CHD absent in their family history and 192 individuals with CHD present.

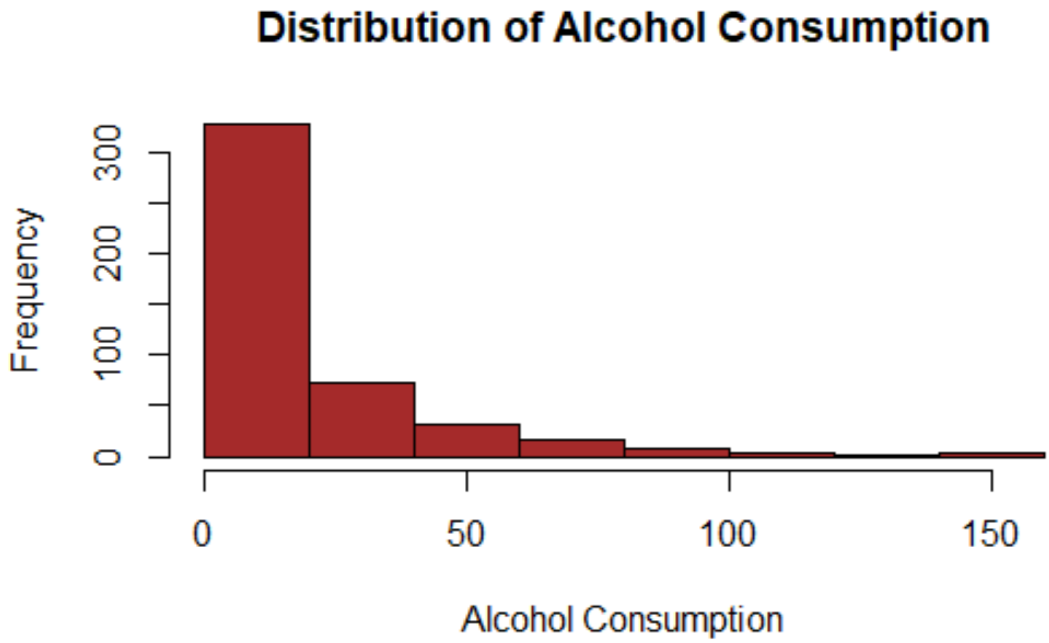

Figure 5. Distribution of alcohol consumption is right skewed with 3 outliers.

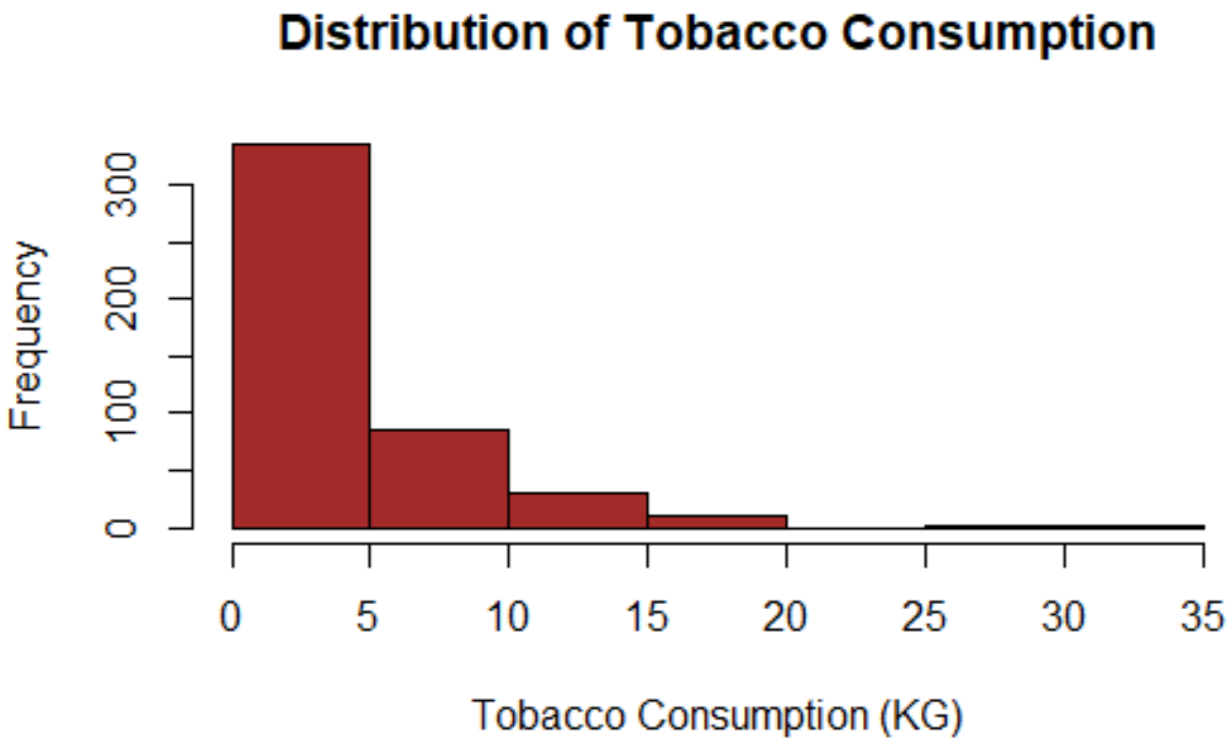

Figure 6. Distribution of tobacco consumption is right skewed with 3 outliers.

Preprocessing:

The preprocessing of SAHeart was relatively brief. It was previously mentioned that CHD was transformed from a numeric variable type to a factor variable type. It was also confirmed that there were no null values in the dataset. Next, the observed outliers were removed from the tobacco and alcohol variables (the top 3 values from each). This concluded the data pre-processing.

Logistic Regression Method:

Logistic regression was the method determined to analyze this data. It is similar to linear regression in that it aims to use an independent variable (or multiple) in order to predict a dependent variable. However, the key difference is that linear regression aims to predict a quantitative result (like height or weight) while logistic regression aims to predict classification

to a category qualitatively (like gender or race). In this case, the model should predict whether a person has CHD based on the other known variables.

Model 1:

The seed is set at 1,234 so that results are reproducible. Without setting, the algorithm would select a random number when using command “sample”. This ensures the training and test datasets always remain the same and are not randomly assigned each time the program is run. An index is made for the data where a 1 and 2 are randomly assigned at a probability of 70% and 30% respectively using the “sample” command. Then, the SAHeart dataset is split into a training dataset (1) and a test dataset (2) based on the number assigned. Thus, each dataset had 348 and 108 rows respectively. Next, the glm (generalized linear model) function was used specifying binomial as the family, causing this function to be a logistic regression. This was generated using the training dataset and aiming to predict presence of CHD using all other variables. The command used to create this model can be seen at the top of Figure 7.

```
> summary(model)

Call:
glm(formula = chd ~ ., family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8045  -0.8394  -0.4577   0.8727   2.4905

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.9264636  1.4848884  -3.991 6.57e-05 ***
sbp             0.0065915  0.0067077   0.983 0.325764
tobacco         0.0713237  0.0337407   2.114 0.034526 *
ldl             0.1812179  0.0673526   2.691 0.007133 **
adiposity       0.0198300  0.0332660   0.596 0.551105
famhistPresent  0.9196491  0.2611536   3.521 0.000429 ***
typea           0.0418479  0.0142574   2.935 0.003334 **
obesity        -0.0745844  0.0484677  -1.539 0.123841
alcohol         0.0006666  0.0062735   0.106 0.915385
age             0.0438685  0.0141086   3.109 0.001875 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 449.62  on 347  degrees of freedom
Residual deviance: 361.32  on 338  degrees of freedom
AIC: 381.32

Number of Fisher Scoring iterations: 5
```

Figure 7. Summary of the logistic regression model generated using the training dataset shows several variables have significant P(z) values including tobacco, LDL, presence in family history, type a behavior, and age.

First, the deviance residuals are a measure of the fit of the model and they “represent the contributions of individual samples to the deviance” (Döring, 2018). Simply put, it is the difference between the predicted value and actual value. If the model is over or undersaturated, the median will not be close to zero. But a median close to zero is one measure to show the model is generally a good fit (Döring, 2018). Since the median is -0.4577 and close to 0, this model is a relatively good fit. Next, there is a list of independent variables in the first column. The first is the intercept which is the value of the dependent variable if all other variables are at 0. The next column, Estimate, lists the coefficients for each value. Coefficients listed give weight to each variable and the number listed is the log odds of contributed by that variable to the total

probability. For example, if family history is present, the coefficient is 0.9196491. Therefore, if family history is present and every other variable remains the same, this factor contributes 0.9196491 to the total log probability that an individual will have CHD. The next significant column is the two-tailed P-value, P(z). A P-value < 0.05 is significant whereas a value > 0.05 is insignificant. This value is the probability that the observed values would occur due to random chance instead of due to correlation. Thus, the lower the P-value, the more probable that the observed values are due to actual relationship (Han et al., 2011). As seen in Figure 7, the asterisks highlight that the following variables are significant: tobacco, LDL, presence of CHD in family history, type a behavior, and age. Since these were the only significant variables, a second model was iterated using only these 5 variables to predict the outcome(Figure 8).

Model 2:

```
> summary(model)

Call:
glm(formula = chd ~ tobacco + ldl + famhist + typea + age, family = binomial,
    data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8897  -0.8330  -0.4570   0.9365   2.5769

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.46201    1.07460  -6.013 1.82e-09 ***
tobacco       0.07086    0.03252   2.179 0.029349 *
ldl           0.16494    0.06168   2.674 0.007492 **
famhistPresent  0.88697    0.25776   3.441 0.000579 ***
typea         0.03981    0.01411   2.821 0.004784 **
age           0.04906    0.01179   4.160 3.19e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 449.62  on 347  degrees of freedom
Residual deviance: 365.14  on 342  degrees of freedom
AIC: 377.14

Number of Fisher Scoring iterations: 4
```

Figure 8. The summary of the second logistic regression model generated using the training dataset was created using only the significant variables from model 1: tobacco, LDL, presence in family history, TABP, and age.

The next important variable in Figure 7/Figure 8 to delineate is the null deviance, also known as the chi-squared value. It is the deviance when all independent variables are 0. If this is low, then that means the data can be modeled simply using the intercept. However, residual deviance is the deviance value when all independent variables are accounted for. Having a residual deviance lower than the null deviance indicates the model is trained well (Döring, 2018).

```
> #output the coefficients and an intercept
> exp(coef(model))
  (Intercept)       tobacco           ldl famhistPresent         typea           age
   0.00156166    1.07342743    1.17931741    2.42776630    1.04060864    1.05028212
```

Figure 9. The coefficients for model 2 show that family history and age have the strongest weight (increasing any of these variables by 1 increases the total probability by the coefficient listed).

The command to show the coefficients after they have been run through the "exp" command should make things clearer. The predictive equation then would look like:

*Y = 1.07342743*X + 1.17931741*W + 2.42776630*V + 1.04060864*U +*

*1.05028212*T + 0.00156166*

In the equation above, X represents the value of tobacco, W represents the value of LDL, and so on. Y represents the log odds of a person having CHD. If all variables are 0, then the probability of a person having CHD is 0.00156166.

```
> model$fitted.values[1:10]
        3         5         6        10        11        12        13        14        15        16
0.33876924 0.68094556 0.64725991 0.67443399 0.56444928 0.66953622 0.04882822 0.03025972 0.55874366 0.19556100
```

Figure 10. The probabilities that a row will be CHD positive using the model to predict based on the given data is fairly accurate.

The fitted values are the probabilities of each row being CHD positive using the model as the prediction mechanism based on the other datapoints. The probabilities for each row can be shown but only the first 10 rows of the training dataset were shown for the sake of

brevity)(Figure 10). Since CHD is a binary category, these probabilities needed to be rounded in order to properly compare them to the actual values. The command seen in Figure 11 was used to round the values and show a confusion matrix which displays the number of values which are correctly and incorrectly identified. The top left value (193) shows the total number of rows which are listed as 0 and were predicted as 0 when the probabilities were rounded. However, the top right value shows the number of values which were predicted as 0 but were actually 1. When the correct values were tallied (top left and bottom right values), they sum to 258 out of 348 total rows. This means the model had a 74% accuracy on the training dataset. The same commands were run for the test dataset (Figure 12) which yielded 82 correctly predicted rows out of 108 yielding a model accuracy of 76% for the test data. It's worth briefly mentioning the residuals plot shows the prediction curve is also logistic in shape (Figure 13). A residual is the observed value minus the predicted value (Han et al., 2011). The sum and mean of all residual values for a given set is 0. In short, since the values of the residual plot are patterned (non-random in shape), it's clear that a linear regression model is not appropriate for the data.

```
> table(round(predict(model, train.data, type="response")), train.data$chd)

    0   1
0 193  56
1  34  65
```

Figure 11. The confusion matrix of the training data shows that 258 rows were correctly predicted based on the probabilities generated by the model while 90 values were incorrectly predicted yielding a 74% accuracy of model 2.

```
> mypredictions<-round(predict (model, test.data, type="response"))
> table (mypredictions, test.data$chd)

mypredictions  0  1
            0 62 15
            1 11 20
```

Figure 12. The confusion matrix of the test data shows that 82 rows were correctly predicted based on the probabilities generated by the model while 26 values were incorrectly predicted yielding a 76% accuracy of model 2.

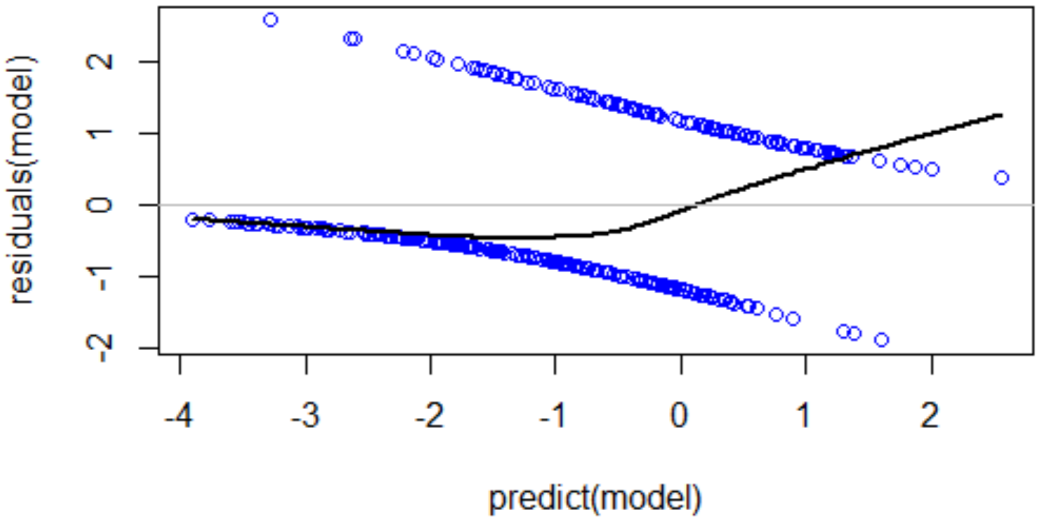

Figure 13. The residuals plot shows the prediction curve is logistic in shape.

Minimal Adequate Model:

A model was also obtained by running the command for the minimal adequate model (MAM)(Figure 14). This iterates through all of the variables to find the model which has the highest accuracy with the least number of variables. In this case, the MAM was the same as model 2 except it added obesity along with the other variables (tobacco, LDL, presence in family history, TABP, and age). So, the only variables not used were SBP, adiposity, and alcohol. The command to create the MAM is seen in Figures 14-15 along with the results of the confusion

matrices for the training and test datasets. This model is nearly equivalent to model 2. The training dataset scored 1 value better, but the test dataset scored 1 value worse.

```
> summary(mamodel)

Call:
glm(formula = chd ~ tobacco + ldl + famhist + typea + obesity +
    age, family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9451  -0.8286  -0.4622   0.9052   2.4842

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.46516    1.24540  -4.388 1.14e-05 ***
tobacco       0.06937    0.03280   2.115 0.034412 *
ldl           0.19261    0.06555   2.938 0.003299 **
famhistPresent  0.90134    0.25926   3.477 0.000508 ***
typea         0.04066    0.01415   2.875 0.004046 **
obesity      -0.04862    0.03225  -1.508 0.131677
age           0.05125    0.01181   4.341 1.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 449.62  on 347  degrees of freedom
Residual deviance: 362.78  on 341  degrees of freedom
AIC: 376.78

Number of Fisher Scoring iterations: 4
```

Figure 14. The summary of the MAM generated using the training dataset shows only obesity is an insignificant variable based on the P-value.

```
> mamodel<-glm(chd~tobacco + ldl + famhist + typea + obesity + age, family=binomial, data=train.data)
> #confusion matrix for the training set
> table(round(predict(mamodel, train.data, type="response")), train.data$chd)

     0   1
  0 196  58
  1  31  63
> #confusion matrix for the test data
> mypredictions<-round(predict (mamodel, test.data, type="response"))
> table (mypredictions, test.data$chd)

mypredictions  0  1
            0 60 14
            1 13 21
```

Figure 15. The confusion matrix of the training data shows that 259 rows were correctly predicted (of 348) yielding a 74% accuracy of the MAM. The confusion matrix of the test data shows that 81 rows were correctly predicted yielding a 75% accuracy of the MAM.

For comparison, the same parameters were run on model 1 (Figure 16). That model yielded a 73% accuracy of the training dataset with a 71% accuracy of the test dataset.

```
> ogmodel<-glm(chd~., family=binomial, data=train.data)
> #confusion matrix for the training set
> table(round(predict(ogmodel, train.data, type="response")), train.data$chd)

     0   1
  0 191  59
  1  36  62
> #confusion matrix for the test data
> mypredictions<-round(predict (ogmodel, test.data, type="response"))
> table (mypredictions, test.data$chd)

mypredictions  0  1
            0 59 17
            1 14 18
```

Figure 16. The confusion matrix of the training data shows that 253 rows were correctly predicted (of 348) yielding a 73% accuracy of model 1. The confusion matrix of the test data shows that 77 rows were correctly predicted (of 108) yielding a 71% accuracy of the MAM.

Results and Model Evaluation:

Thus, 3 models were made: model 1 (using all variables to predict CHD); model 2 (using only the significant variables to predict CHD); and MAM (using the same variables as model 2 plus obesity). The MAM showed a 3% increase in accuracy over model 1 and model 2 had a 1% increase over the MAM. As previously mentioned, all variables listed except for alcohol are known to be risk factors in CHD. Adiposity, obesity, SBP, and alcohol were the only factors not deemed significant based on Figure 7. Interestingly, there were some false negatives observed along with true negatives. Alcohol is shown by the literature not to have a negative impact on health towards CHD (unless taken in excess amounts). If anything, there is a negative relationship observed between the two (Marmot, 2001). Therefore, the model accurately handled the alcohol variable based on what is known in the data. On the other hand, TABP may or may not be a causal factor when it comes to CHD (Baker et al., 2007). Since the dataset is regarding a

South African population, it could be that TABP is a causal factor for CHD for that particular population whereas it is not for other populations. Thus, it cannot be said with 100% certainty whether the model accurately handled the TABP variable because the literature is split regarding its effects.

On the other ends of the spectrum, adiposity, obesity, and SBP are all known and well-characterized as having a negative impact towards CHD. Yet, the model did not find that these significantly contributed to CHD. It could be that the dataset was too small and given a larger dataset the P-values would decrease to show these variables were significant. In this case, it is important to point out that the highest P-value of these 3 was 0.5 – this is not a very high value meaning that it is possible this could decrease significantly given more datapoints. It could also be that this particular South African population sees less contribution towards CHD as a result of obesity, adiposity, and/or SBP. This is unlikely - SBP is a measure which indicates stress on the cardiac system suggesting CHD whereas high obesity/adiposity levels are the direct result of unhealthy diet resulting in higher LDL levels and plaque in the cardiac arteries. Thus, it is more likely that the observed results are due to random chance in a smaller dataset and these variables would be seen as significant in a larger dataset.

When it comes to the actual accuracy of the models, it is clear to see how MAM and model 2 were more accurate than model 1. But it is also important to note that having no model means one would only be 50% accurate (since having CHD is a binary variable – present/absent). Thus, using the MAM or model 2 gives at least a 25% increase in accuracy. It is expected that having an expanded dataset would help predict at a higher rate. Normally, it would be expected that the algorithm to create the MAM would be precise in its iteration and generate a model more accurate than could be custom made. However, it was shown that model 2 was

slightly more accurate (by 1%) and this was only due to the MAM having the extra variable of obesity. It is also worth mentioned briefly that the null deviance for all 3 models was greater than the residual deviance and that the null deviance was not 0. This means that the data was being predicted using the models better than without the models.

When examining the metrics of the variables themselves, age was by far the best predictor with a P-value in the 10^{-5} range for model 2 and the MAM. For model 2, there was a 0.000319% chance that the data would appear in the observed configuration due to random chance. Subsequently, family history was the second best with a P-value in the 10^{-4} range. The successive best predictors were following these by a magnitude of 10 and 100. These two being the best predictors make intuitive sense. CHD naturally happens over time. Even if one takes strong, specific, purposeful measures to prevent plaque from building, it still possible it will build up given enough time. On the other hand, family history is also a great predictor likely because of the combination between genetic predisposition to CHD and natural socialization into behaviors associated with CHD (smoking, sedentary lifestyle, unhealthy diet, etc.).

Conclusion:

It has thus been shown that according to the SAHeart dataset, LDL levels, TABP, age, tobacco usage, and knowing if a person has a family history of CHD can be used to predict CHD in a person at a 75% accuracy. Three models were generated in order to predict whether a person had CHD: model 1 (all variables); model 2 (only significant variables); Minimal Adequate Model (model generated by MAM algorithm). Model 2 and the MAM were shown to be the best. Age was by far the best predictor followed by if a person has a family history of CHD.

Limitations and Improvements:

There are several limitations this study saw and a few improvements which could be taken on for future work. First, nearly every variable in the dataset is known to contribute to CHD (with the exception of alcohol). While theoretically the contributing variables analyzed are more interesting, it is necessary to have variables which act as a negative control. Ideally, there would be more than one to ensure the integrity of the data and analyses. Second, this data did not capture any exercise metrics which is one of the major interventions for CHD. Having how much a person exercises per week over their lifetime would be invaluable to the analysis. In addition, the other risk factors previously discussion should also be collected for a new dataset: all types of cholesterol (HDL, triglycerides, and total), diabetes status, and diet. It was previously mentioned this dataset may be too small since obesity, adiposity, and SBP were all shown to not be significant. An expanded dataset should be obtained with a larger sample size in order to determine if these results truly are not significant to this population or if they just were not found to be significant due to the dataset size.

References:

Baker, G. J., Suchday, S., & Krantz, D. S. (2007). Heart disease/attack.

C.D.C. (2020). *Heart Disease Facts*. Center for Disease Control. Retrieved from: https://www.cdc.gov/heartdisease/facts.htm

Döring, M. (2018). Interpreting Generalized Linear Models. Data Science Blog. Retrieved from: https://www.datascienceblog.net/post/machine-learning/interpreting_generalized_linear_models/

Han, J., Kamber, M., and Pei, J. (2011). Data Mining: Concepts and Techniques, Third Edition. Elsevier. Retrieved June 6th, 2021 from:

http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

KEEL. (n.d.). *South African Heart Dataset*. Soft Computing and Intelligent Information Systems. Retrieved from: https://sci2s.ugr.es/keel/dataset.php?cod=184

Marmot, M. G. (2001). Alcohol and coronary heart disease. *International Journal of Epidemiology*, *30*(4), 724-729.

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., ... & Turner, M. B. (2015). Heart disease and stroke statistics—2015 update: a report from the American Heart Association. *circulation*, *131*(4), e29-e322.

Petticrew, M. P., Lee, K., & McKee, M. (2012). Type A behavior pattern and coronary heart disease: Philip Morris's "crown jewel". *American journal of public health*, *102*(11), 2018–2025. https://doi.org/10.2105/AJPH.2012.300816

Rosenman, R. H., & Chesney, M. A. (1980). The relationship of type A behavior pattern to coronary heart disease. *Activitas Nervosa Superior*, *22*(1), 1-45.

Schwalbe F. C. (1990). Relationship between Type A personality and coronary heart disease. Analysis of five cohort studies. *The Journal of the Florida Medical Association*, *77*(9), 803–805.

Torpy, J. M., Burke, A. E., & Glass, R. M. (2009). Coronary heart disease risk factors. *Jama*, *302*(21), 2388-2388.