Table of Contents

1.0 Introduction

1.1 Background

Coronary Heart Disease (CHD), also known as coronary artery disease (CAD), is a cardiovascular condition that primarily affects the blood vessels supplying the heart muscle. It is characterized by the narrowing, blockage, or hardening of the coronary arteries, which are responsible for delivering oxygen-rich blood to the heart (Centers for Disease Control and Prevention, 2021).

Coronary Heart Disease is influenced by several risk factors, both modifiable and non-modifiable. Modifiable risk factors include unhealthy lifestyle habits such as smoking, poor diet high in saturated fats and cholesterol, physical inactivity, and obesity. Non-modifiable risk factors include age, family history of heart disease, gender (men are at higher risk than pre-menopausal women), and ethnicity (some ethnic groups are at higher risk) (Hajar, 2017).

CHD is a significant global health problem and a leading cause of death in many countries. It affects millions of people worldwide and places a substantial burden on healthcare systems (Cardiol, 2010). Despite its seriousness, CHD is largely preventable and can be managed effectively through lifestyle modifications, early detection, and appropriate medical interventions.

Preventive measures, such as adopting a heart-healthy diet, engaging in regular physical activity, maintaining a healthy weight, managing blood pressure and cholesterol levels, avoiding tobacco use, and managing diabetes, can significantly reduce the risk of developing Coronary Heart Disease. Additionally, early detection and timely medical intervention, including the use of medications, angioplasty, stent placement, or coronary artery bypass surgery, when necessary, can help manage the condition and improve outcomes for individuals with CHD.

1.2 Overview

We have selected the Coronary Heart Disease dataset from Kaggle for our analysis. This dataset comprises data from 463 individuals. It includes various columns related to coronary heart disease, enabling us to assess its prevalence for each person. The dataset contains systolic blood pressure, yearly tobacco use (in kg), low-density lipoprotein (ldl), family history (0 or 1), type A personality score (typea), obesity (body mass index), alcohol use, age, and the diagnosis of CHD (0 or 1). To perform our analysis, we consider the factors affecting coronary heart disease as the regressor variables, denoted as "x." On the other hand, the coronary heart disease will be treated as the dichotomous response variable, denoted as "y." This allows us to explore the relationship between the various factors and the occurrence of coronary heart disease in the dataset.

y : Coronary heart disease
x_1 : Systolic blood pressure
x_2 : Yearly tobacco use (in kg)
x_3 : Low density lipoprotein (ldl)
x_4: Family history (0 or 1)
x_5 : Type A personality score (typea)
x_6: Obesity (body mass index)
x_7: Alcohol use
x_8: Age

Multiple regression analysis is utilized when there are multiple regressors (>1) to model and describe the relationships among variables in a dataset.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$

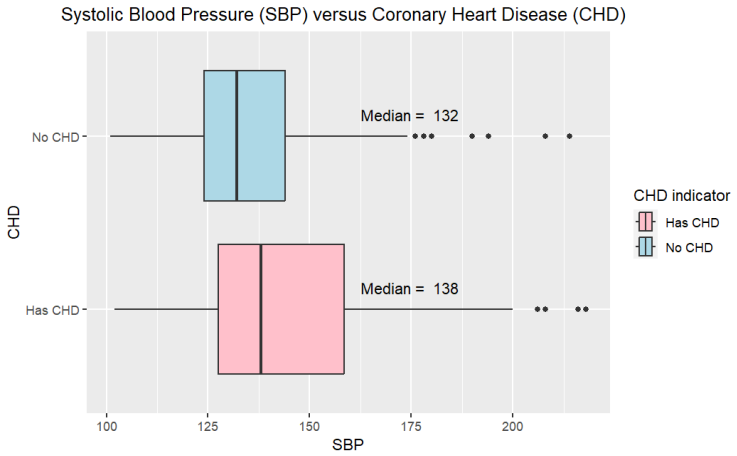
1.3 Objective

The primary aim of this study is to explore the associations between different risk factors and coronary heart disease (CHD) using multiple regression analysis and other relevant statistical techniques on a comprehensive dataset. Our specific goal is to identify the significant predictors that exert a substantial influence on the probability of developing coronary heart disease. Through this analysis, we seek to gain valuable insights into the key risk factors that contribute significantly to the occurrence of CHD and potentially improve risk assessment and preventive strategies for this cardiovascular condition.

2.0 Graph Visualisation

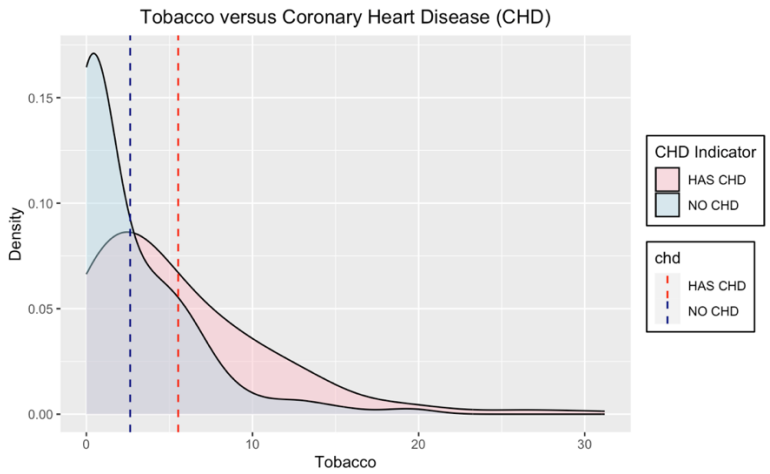
2.1 Box Plot, Cumulative Density Plots & Logistics Regression Plot

(a) Systolic Blood Pressure (SBP) versus Coronary Heart Disease (CHD)



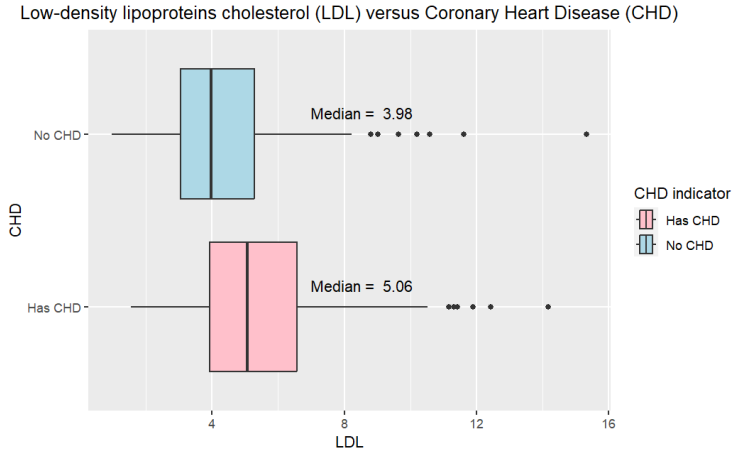
Based on the plot above, both batches of data appear to be right skew as the median line is towards the left of the box. The median of individuals with CHD due to systolic blood pressure (SBP) is 138 which is higher than that of those without coronary heart disease with a median of 132. The gap between the median lines of the two groups has a decent amount. Hence, we can suggest that SBP regressor is slightly significant to CHD.

(b) Tobacco use versus Coronary Heart Disease (CHD)



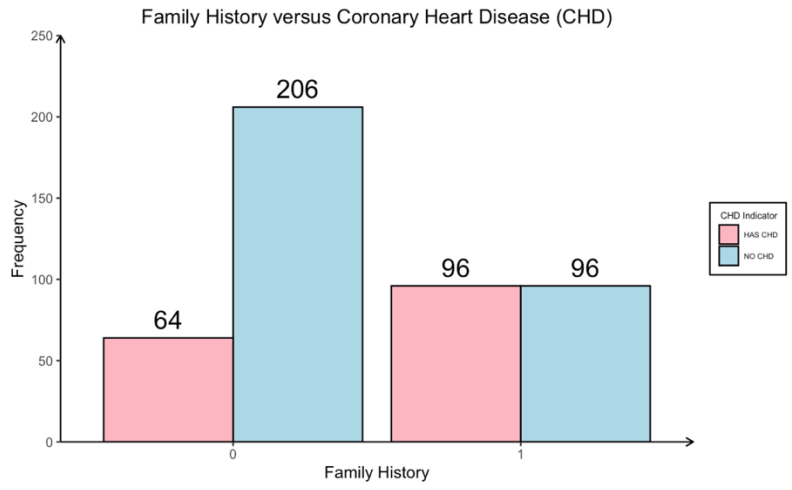
The density graph above clearly indicates that the density plot on the presence of coronary heart disease (CHD) is noticeably different from the absence of coronary heart disease (CHD) based on the Tobacco. According to scientific finding, high tobacco uses such as cigarette smoking is the major cause of coronary heart disease (Centers for Disease Control and Prevention, 2010). Hence, this striking contrast suggested that Tobacco plays a highly significant effect in determining the existence of coronary heart disease. We can suggest that tobacco use is a significant factor on coronary heart disease based on the visualisation of graph.

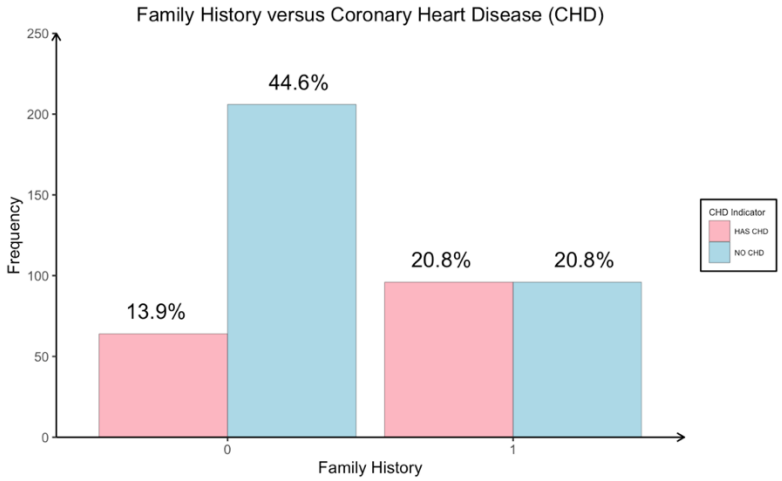
(c) Low-density lipoproteins cholesterol (LDL) versus Coronary Heart Disease (CHD)



Based on the plot above, both batches of data appear to be slightly right skew as the median line is towards the left of the box. Low density lipoproteins, often referred to as “bad cholesterol” poses a risk when its levels are elevated in the bloodstream. High low-density lipoproteins (LDL) cholesterol can lead to the narrowing and blockage of arteries, potentially resulting in heart problems (WebMD, 2020). The median of individuals with coronary heart disease (CHD) due to LDL is 5.06 which is higher than that of those without coronary heart disease with the median of 3.98. Besides that, the gap between the median lines of the two groups is large. This suggests that a high level of LDL increases the risk for CHD and thus, suggests that LDL is statistically significant to CHD.

(d) Family History versus Coronary Heart Disease (CHD)

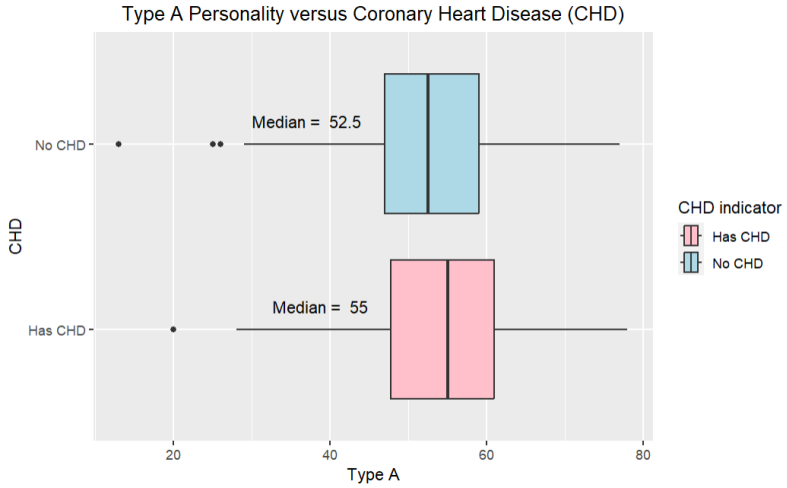




Let "0" denotes as "Absent" of family history and "1" denotes as "Present" of family history.

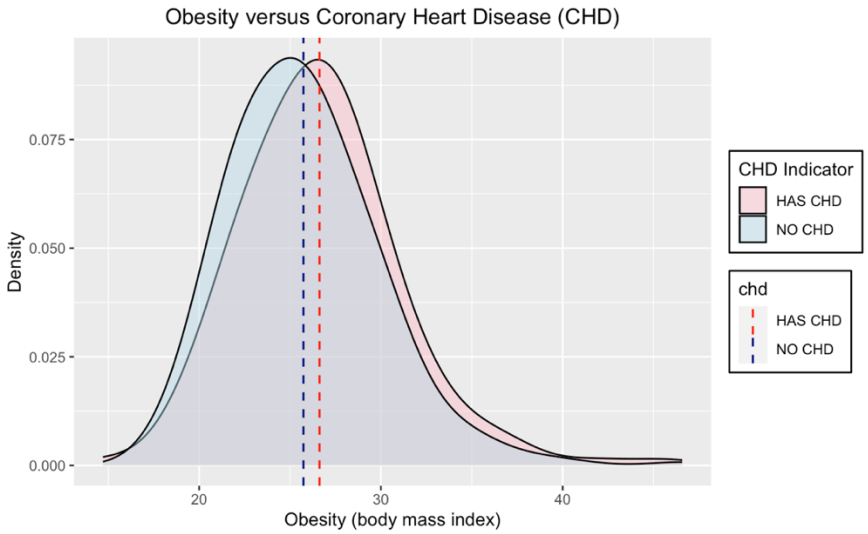
Based on the scientific research, we found that an individual is more likely to develop coronary heart disease if he or she have a family health history of heart disease (Centers for Disease Control and Prevention, 2022). Based on the bar chart above, we noticed if family history background existed (value of 1), the number of individuals in presence and absence of coronary heart disease is the same to each other which is 96 individuals (20.8%) from the data set. On the other hand, when there's no family history background of disease, there is lower chance to get suffer in coronary disease (13.9%) as shown in the bar chart above. Hence, based on the bar chart, we can suggest that the existence of family history background will substantially affect the occurrence of coronary heart disease.

(e) Type A Personality versus Coronary Heart Disease (CHD)



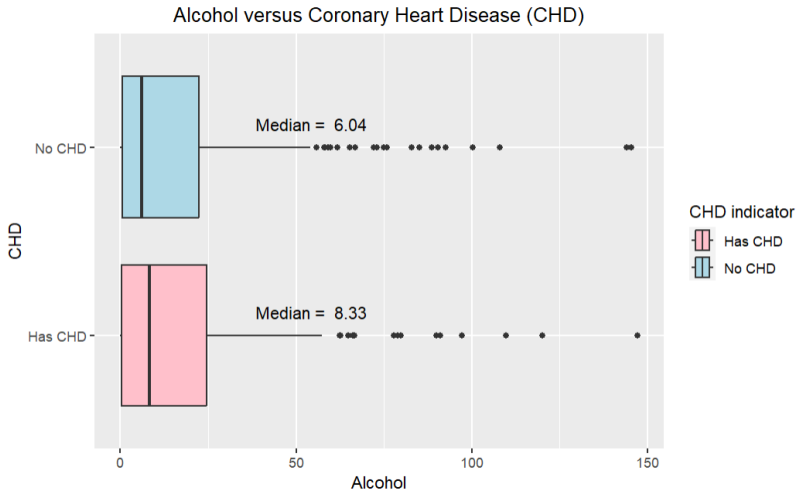
The data of individuals with coronary heart disease (CHD) appear to be left skew as the median line is towards the right of the box whereas the data of individuals without CHD appear to be right skew as the line is towards the left of the box. The median of individuals with CHD due to Type A personality is 55 which is higher than that of those without CHD with the median of 52.5. The plot above shows that the gap between the median lines of the two groups is small. This suggests that Type A personality is an insignificant regressor to our model.

(f) Obesity versus Coronary Heart Disease (CHD)



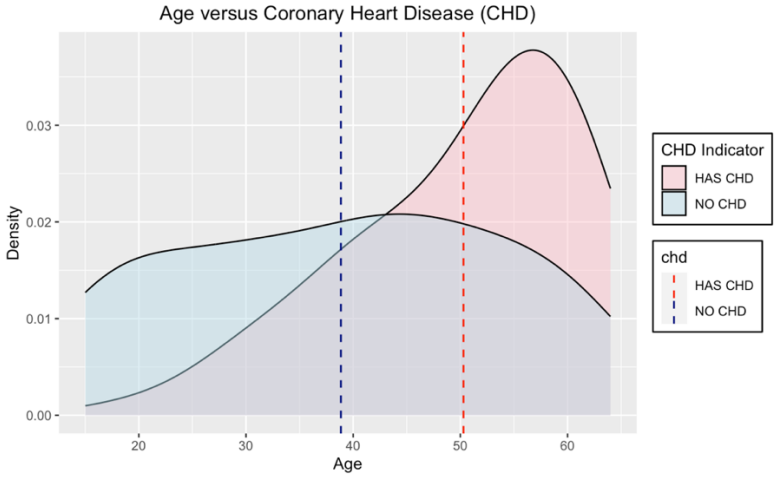
In theoretical speaking, obesity in term of body mass index (BMI) is often considered as relatively minor coronary heart disease (CHD) risk factor (Ades & Savage, 2017). Based on the density graph above, it is clearly implied that both density graph of "Has CHD" and "No CHD" versus obesity is approximately similar and the mean line of both response categorical data is very close to each other. Owing to the similar graph trend and the mean, it is suggested that the obesity factor has no significant effect on the existence of coronary heart disease based on the dataset. Note that this might be due to the average BMI of the participants in this dataset is considerably low, where there are less individuals having high BMI (30 and above).

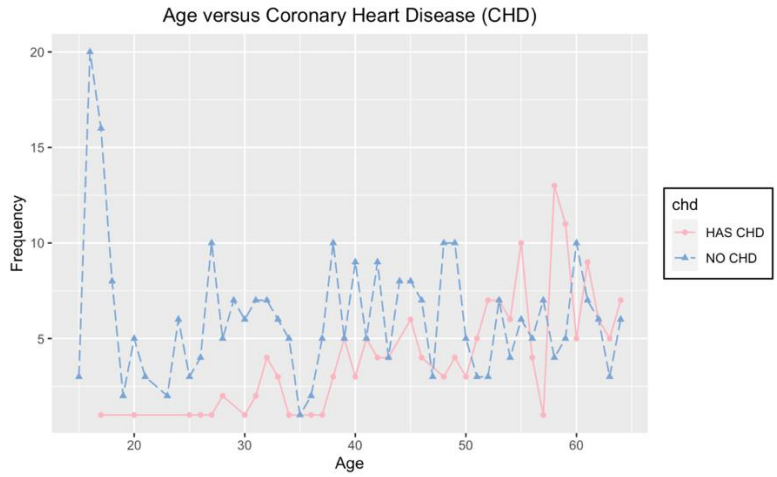
(g) Alcohol versus Coronary Heart Disease (CHD)



Both batches of data appear to be right skew as the median line is towards the left of the box. The median of individuals with coronary heart disease (CHD) due to alcohol is 8.33 which is higher than that of those without coronary heart disease with the median of 6.04. Theoretically, drinking too much alcohol can raise blood pressure levels and the risk for heart disease (Centers for Disease Control and Prevention, 2019), but since our data shows that the alcohol consumption of most individuals is low, it might have translated to a small gap between the median lines of the two groups. Additionally, certain studies have indicated that moderate drinking, which translates to one drink per day for women and two drinks per day for men, is associated with reduced risks of mortality from heart disease (Web MD, n.d.). Thus, for our model, we can suggest that alcohol is statistically insignificant to CHD.

(h) Age versus Coronary Heart Disease (CHD)





The density graph above implies the relationship between the age and the existence of coronary heart disease (CHD). According to the finding, Age plays a crucial role in the deterioration of cardiovascular functionality, resulting in an increased risk of coronary heart disease in older adults (Jousilahti et.al., 1999). Based on the graph above, the spread of the density graph between age and existence of coronary heart disease indicates a great contrast. Besides, there indicates a great difference between both means. Hence, there are substantial evidence to suggest that the age regressor is highly significant to the existence of coronary heart disease.

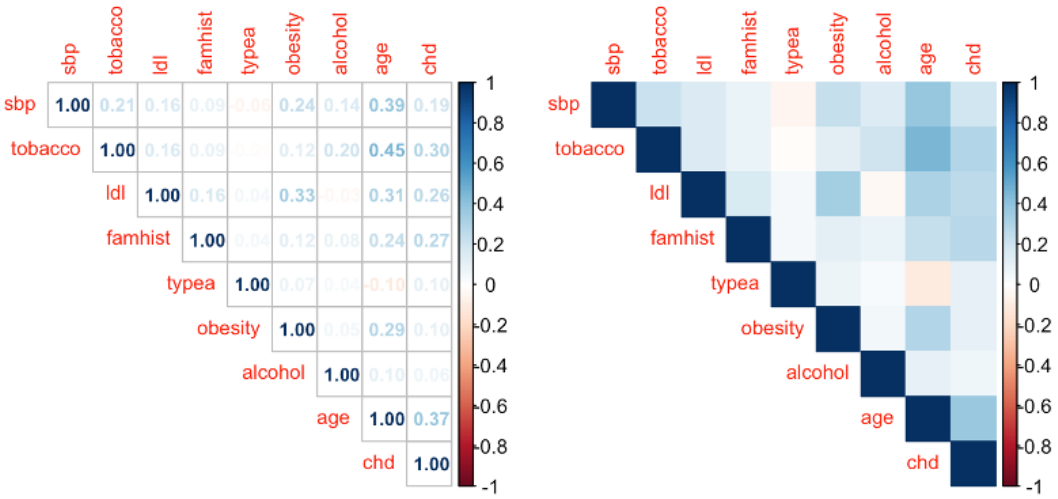
The line graph above caters the information of individuals from age 15 to 64 that suffers from coronary heart disease (CHD). By referring to the line graph, most of the individuals suffer in coronary heart disease at age 58. An overall of positive trend is shown by the presence of coronary heart disease (pink trendline). As an individual become older, the number of individuals suffering from coronary heart disease is higher.

2.2 Correlation between Regressor Variables

In our data of cases of coronary heart disease, we aimed to investigate the correlation between coronary heart disease and risk factors. The two variables of interest are the dichotomous response variable "Coronary Heart Disease" and "Risk Factors" (with regressor variables "Systolic Blood Pressure", "Yearly Tobacco Use", "Low Density Lipoprotein", "Family History", "Type A Personality Score", "Obesity", "Alcohol Use" and "Age").

Additionally, to assess the strength of association between these two categorical variables, we employed the Phi coefficient, a measure designed for contingency tables of categorical data. The Phi coefficient is a statistic that ranges between -1 and 1, where -1 represents a perfect negative association, 0 indicates no association, and 1 indicates a perfect positive association.

After analysing the data, we found the following correlation values:



(a) Coronary Heart Disease versus Systolic Blood Pressure:

With a correlation coefficient of 0.1923541, we can infer that there is a weak positive relationship between coronary heart disease and systolic blood pressure. This means that as systolic blood pressure increases, there is a slight tendency for the occurrence of coronary heart disease to increase as well.

(b) Coronary Heart Disease versus Yearly Tobacco Use:

With a correlation coefficient of 0.2997175, we can deduce that there exists a mild to moderate positive correlation between coronary heart disease and yearly tobacco use. This indicates that, on average, when yearly tobacco use increases, there is a mild tendency for the occurrence of coronary heart disease to also exhibit a rise.

(c) Coronary Heart Disease versus Low Density Lipoprotein:

Given a correlation coefficient of 0.2630527, we can conclude that there is a mild to moderate positive correlation between coronary heart disease and Low-Density Lipoprotein. This means that there is a mild tendency where elevated LDL cholesterol levels can contribute to the development of atherosclerosis and subsequent CHD.

(d) Coronary Heart Disease versus Family History:

With a correlation coefficient of 0.2723727, we can infer that there is a mild to moderate positive correlation between coronary heart disease and family history. This means that if a person has one or more first-degree relatives (parents or siblings) who have experienced CHD at a young age (usually before age 55 for men and 60 for women), there is a mild tendency that their risk may be higher (Chow, 2007).

(e) Coronary Heart Disease versus Type A Personality Score:

With a correlation coefficient of 0.1031558, we can infer that there is a very weak positive correlation between coronary heart disease and Type A personality score. This means that on average, there is a minor tendency for individuals with higher Type A personality scores to have a slightly higher prevalence of coronary heart disease.

(f) Coronary Heart Disease versus Obesity:

Having a correlation coefficient of 0.1000951, we can conclude that there exists a very weak positive correlation between coronary heart disease and obesity. This means that obese people have a minor tendency to have higher likelihood of developing CHD compared to those with a normal weight.

(g) Coronary Heart Disease versus Alcohol Use:

With a correlation coefficient of 0.06253068, we can deduce that there exists a very weak positive correlation between coronary heart disease and alcohol use. This indicates that there is a minor tendency for high alcohol consumption individuals to have a higher chance of contracting CHD.

(h) Coronary Heart Disease versus Age:

With a correlation coefficient of 0.379733, we can conclude that there exists a moderate positive correlation between coronary heart disease and age. This means that an older age will lead to the highest tendency of contracting CHD compared to the other regressor variables.

2.3 Data summaries

Regressors	Correlation
Systolic Blood Pressure (SBP)	0.1923541
Yearly Tobacco Use	0.2997175
Low-density Lipoprotein (LDL)	0.2630527
Family History	0.2723727
Type A Personality	0.1031558
Obesity	0.1000951
Alcohol Use	0.06253068
Age	0.3729733

Note that a high correlation value between variables does not necessarily indicate that the regressor variable (predictor variable) is significant in a regression analysis. Correlation measures the strength and direction of the linear relationship between two variables, but it does not provide information about the statistical significance or causal relationship between them. Nonetheless, a high correlation value does suggest that changes in the predictor variable (regressor variable) will associate to a significant change in response variable. Similarly, we can say a high correlation value could translate to a high regressor coefficient value (β).

3.0 Generalized Linear Models (GLM)

3.1 Appropriateness of Binomial Logistics Regression as GLM Model

In the realm of statistical analysis, binomial logistic regression stands as a powerful tool for investigating the relationship between a binary response variable and one or more predictor variables. However, before proceeding with our analysis, it is essential to ensure that our data meets the fundamental assumptions necessary for a valid and reliable logistic regression model. Hence, we will demonstrate how our data fulfils six key assumptions.

3.1.1 Assumptions to be Met

Assumption 1: Dichotomous Dependent Variable

Our first step in assessing our data was to examine the dependent variable, which should consist of two categorical, independent groups, also known as a dichotomous variable. Similarly, it is also known as a binomial variable. Our data adheres to this criterion as it comprises two mutually exclusive and exhaustive categories which is “No CHD” and “Has CHD”.

Assumption 2: Independent Variables at the Continuous or Nominal Level

Another crucial aspect of our data assessment involved evaluating the independent variables. Our data indeed consists of two or more independent variables, all measured either at the continuous or nominal level. Continuous variables, such as "Systolic Blood Pressure", "Yearly Tobacco Use", "Low Density Lipoprotein", "Type A Personality Score", "Obesity", "Alcohol Use" and "Age" were measured with precision. Note that “Family History” is the only non-continuous regressor variable.

Assumption 3: Independence of Observations

To proceed with logistic regression, we had to verify the independence of observations. This means that there should be no relationship between the observations, and each data point should stand as an independent entity. For simplicity’s sake, we will assume an independence of observation. This is because there is no easy method to verify datasets obtained from Keggle.

Assumption 4: Absence of Multicollinearity

In our analysis, the Variation Inflation Factors (VIF) will be used to confirm the absence of significant multicollinearity. Basically, a VIF value greater than 5 warrants further investigation while a VIF value greater than 10 indicates serious multicollinearity.

R code to conduct VIF Test for full GLM model:

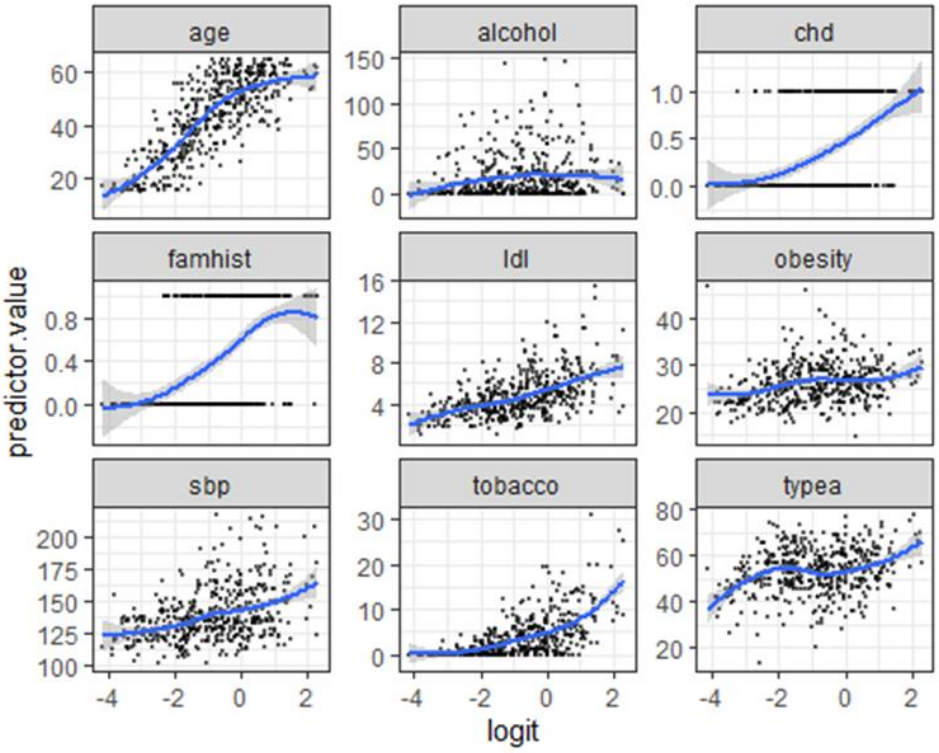
```
###### VIF Test for Multicollinearity ######
car::vif(fullModel)
```

Results obtained from VIF Test:

Regressors	**VIF Value**
Systolic Blood Pressure (x_1)	1.146515
Usage of Tobacco (x_2)	1.164828
Low-density lipoproteins cholesterol (x_3)	1.132787
Family History (x_4)	1.020699
Type A Personality (x_5)	1.082978
Obesity (x_6)	1.159339
Alcohol intake (x_7)	1.064402
Age (x_8)	1.319792

According to the VIF results obtained above, all 8 regressors are having an obviously small VIF value, which is below 2. Hence, we can conclude that all the regressors are orthogonal to each other and there is no correlation between a particular regressor with other regressors.

Assumption 5: Linear Relationship between Continuous Independent Variables and the Logit Transformation



In our regression analysis, we examined scatter plots between each continuous independent variable and the logit transformation of the dependent variable. Upon visual inspection, we observed that all the scatter plots displayed a relatively straight line or a clear linear pattern. This visual evidence strongly suggests a linear relationship between the continuous independent variables and the logit transformation of the dependent variable. Note that graphs of CHD and Family History should be disregarded because they are not continuous independent variables. They are plotted together through a R package code.

As additional information, a steeper best-fit line would indicate a significant regressor while a flat best-fit line would indicate otherwise. Thus, we notice regressor variables for alcohol, obesity and Systolic Blood Pressure display a flat best-fit line. We shall conduct further analysis to verify this hypothesis in Content 3.2.

Assumption 6: Absence of Significant Outliers, High Leverage Points, and Highly Influential Points

In our regression analysis, we have made the decision to proceed with the complete dataset, fully aware that it may include outliers, leverage points, and highly influential points. Despite the presence of these potentially unusual observations, we have chosen to work with the entire dataset at this stage of the analysis. By doing so, we aim to capture the full range of variability and complexities present in the data, allowing for a comprehensive exploration of the relationships between the independent and dependent variables.

3.2 Full GLM Model

The summary statistics from the full fitted GLM model will be shown as below:

```
Call:
glm(formula = chd ~ ., family = binomial(link = "logit"), data = CoronaryHeart_Disease)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.4169865  1.2401631  -5.174 2.29e-07 ***
sbp          0.0067329  0.0057297   1.175  0.23996
tobacco      0.0795655  0.0266441   2.986  0.00282 **
ldl          0.1824114  0.0582845   3.130  0.00175 **
famhist      0.9234083  0.2276771   4.056 5.00e-05 ***
typea        0.0389489  0.0122687   3.175  0.00150 **
obesity     -0.0422070  0.0294398  -1.434  0.15167
alcohol      0.0002806  0.0044809   0.063  0.95006
age          0.0489856  0.0106015   4.621 3.83e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.55  on 453  degrees of freedom
AIC: 490.55

Number of Fisher Scoring iterations: 5
```

The fitted model equation is stated as below:

$$y = -6.4169865 + 0.0067329x_1 + 0.0795655x_2 + 0.1824114x_3 + 0.9234083x_4 + 0.0389489x_5 - 0.0422070x_6 + 0.0002806x_7 + 0.0489856x_8$$

3.2.1 Significance of Regressors Variables

Confidence Intervals for Coefficients

Coefficient	2.5%	97.5%
Intercept	-8.914892014	-4.040931535
SBP (x_1)	-0.004451461	0.018093192
Tobacco (x_2)	0.028621275	0.133424103
LDL (x_3)	0.070410323	0.299718368
Family History (x_4)	0.479494804	1.373327762
Type A Personality (x_5)	0.015320624	0.063521838
Obesity (x_6)	-0.101165997	0.014602711
Alcohol (x_7)	-0.008625906	0.009062124
Age (x_8)	0.028570592	0.070239093

Confidence intervals can be used to determine the significance levels of each coefficient. Based on the summary above, the confidence intervals for coefficients includes two columns, one representing the upper confidence limit and the other representing the lower confidence limit for each parameter. For variables x_1, x_6 and x_7, we can observe that their confidence intervals include zero as their lower confidence limits are negative and upper confidence limits are positive. When the confidence intervals include zero, there is a likelihood that the regressor coefficient may take a value of 0. In this case, the regressor variable will have no effect on the value of the response variable, deeming it insignificant. Thus, we can conclude that variables x_1, x_6 and x_7 are statistically insignificant at the significance level of 95%.

Pearson's Chi-squared Test

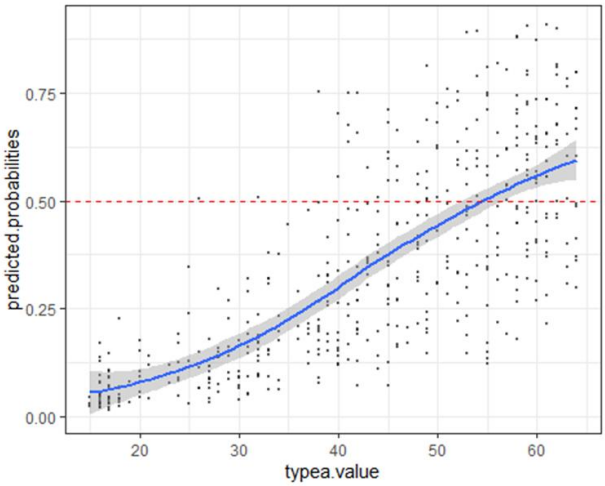
```
        Pearson's Chi-squared test with Yates' continuity
        correction

data:  Assignment$famhist and Assignment$chd
X-squared = 33.123, df = 1, p-value = 8.653e-09
```

As x_4 is the only categorical variable among all the coefficients, additional testing is required to assess its significance level accurately. The chi-square test of independence is used to test whether two categorical variables are related to each other (Turney, 2022). Based on the result above, we can observe that the p-value is smaller than 0.05. This indicates that the family history is statistically significant to the CHD.

Further Investigation on variable x_1 and x_5

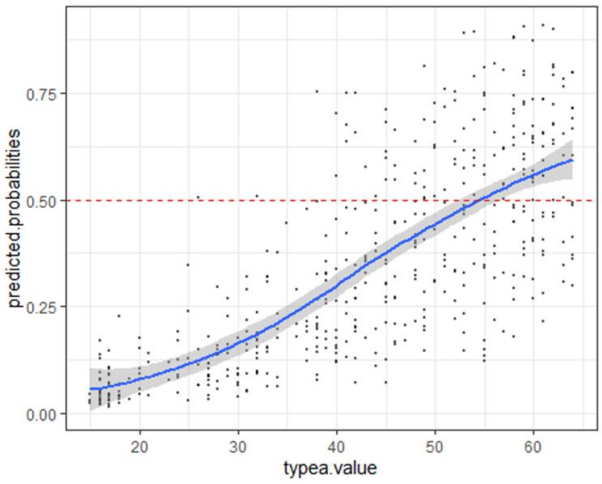
Based on the GLM model's summary statistics, it is shown that "Systolic Blood Pressure", 'Type A Personality" and "Obesity" regressor variables are insignificant. This summary statistics creates 2 contradictions from our suggestions based on our Data Visualization. To justify, the significance of regressor variables "Systolic Blood Pressure" and "Type A Personality" were wrongly suggested through Data Visualization in Content 2.0. Thus, we will apply extra graphical analysis using binomial logistic regression with predicted probabilities to visualize the significance of these regressors.

(a) Systolic Blood Pressure (SBP)



To further investigate the significance level of SBP, we produce a scatter plot of predicted probabilities with its SPB value. A trendline is added to the plot showing the best fit to the data. Based on the plot above, we notice that the data points are randomly distributed across the entire space. If the data points on the scatter plot appear to be distributed randomly, it indicates that there is no discernible relationship between the variables (CK12-Foundation, n.d.). Thus, through extra verification of visualization, we can conclude that SBP is statistically insignificant to CHD.

(b) Type A Personality



To further investigate the significance level of Type A, we produce a scatter plot of predicted probabilities with its typea value. A trendline is added to the plot showing the best fit to the data. Based on the scatterplot, we notice that the data points display a upward pattern across the entire space. It can also be noted that there consist a considerable amount of range in the beginning where predicted probabilities are below 0.5. Since there is a specific pattern observed in the plot of data points, Type A personality is said to be significant to CHD. According to Ronesh Sinha, M.D., a Palo Alto Medical Foundation internal medicine doctor, individuals with a Type A personality, characterized by impatience, aggression, and strong competitiveness, are at a greater risk of developing heart disease. Thus, through extra verification of visualization, we can conclude that Type A personality is statistically significant to CHD.

4.0 Reduced GLM

Method 1 : Manual Stepwise Selection

Based on the result of previous testing, we produce a new GLM by manually eliminating the insignificant variables x_1, x_6 and x_7 accordingly. We will conduct further investigation to determine whether it is necessary to remove the regressors by observing the AIC value of the reduced GLM.

Case 1 : Assume x_7 is insignificant

```
Call:
glm(formula = chd ~ sbp + tobacco + ldl + famhist + typea + obesity +
    age, family = binomial(link = "logit"), data = Assignment)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.416927   1.240101  -5.175 2.28e-07 ***
sbp             0.006780   0.005683   1.193  0.23286
tobacco         0.079886   0.026157   3.054  0.00226 **
ldl             0.182102   0.058077   3.136  0.00172 **
famhistPresent  0.924464   0.227061   4.071 4.67e-05 ***
typea           0.038966   0.012266   3.177  0.00149 **
obesity        -0.042200   0.029437  -1.434  0.15169
age             0.048927   0.010556   4.635 3.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.55  on 454  degrees of freedom
AIC: 488.55
```

Case 2 : Assume x_1 and x_7 are insignificant

```
Call:
glm(formula = chd ~ tobacco + ldl + famhist + typea + obesity +
    age, family = binomial(link = "logit"), data = Assignment)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.70273    1.07640  -5.298 1.17e-07 ***
tobacco         0.07999    0.02598   3.079  0.00208 **
ldl             0.18372    0.05818   3.158  0.00159 **
famhistPresent  0.91610    0.22645   4.046 5.22e-05 ***
typea           0.03827    0.01222   3.133  0.00173 **
obesity        -0.03760    0.02910  -1.292  0.19638
age             0.05211    0.01024   5.087 3.63e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 473.98  on 455  degrees of freedom
AIC: 487.98

Number of Fisher Scoring iterations: 5
```

Case 3 : Assume x_1, x_6 and x_7 are insignificant

```
Call:
glm(formula = chd ~ tobacco + ldl + famhist + typea + age, family = binomial(link = "logit"),
    data = Assignment)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.44644    0.92087  -7.000 2.55e-12 ***
tobacco         0.08038    0.02588   3.106  0.00190 **
ldl             0.16199    0.05497   2.947  0.00321 **
famhistPresent  0.90818    0.22576   4.023 5.75e-05 ***
typea           0.03712    0.01217   3.051  0.00228 **
age             0.05046    0.01021   4.944 7.65e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 475.69  on 456  degrees of freedom
AIC: 487.69

Number of Fisher Scoring iterations: 5
```

Based on the above data outputs, we can observe that the AIC value decreases from 490.55 to 488.55 when the first regressor, x_7 is eliminated. After eliminating the second regressor, x_1, the AIC value dropped again from 488.55 to 487.98. When the third regressor, x_6 is removed, we obtained a decrement in AIC value from 487.98 to 487.69. The sequence of elimination is based on the p-value, higher p-value will prior to be eliminated. After three rounds of backward stepwise elimination, we obtained a new AIC value which is 487.69. This is obviously lesser than the original GLM's AIC value which is 490.55. AIC is a metric that is used to compare the fit of different regression models. The model with the lowest AIC offers the best fit (Zach, 2021). Thus, a decrement of 2.86 indicates that the elimination of these three insignificant variables is truly effective for our model.

4.1 Measures of Fit (Verification of the best model)

4.1.1 Penalized Loglikelihood Tests

(a) Akaike Information Criterion (AIC)

The penalty is 2 for each parameter.

$AIC = -2l + 2p$, where l is the loglikelihood and p is the number of parameters estimated $(k + 1)$

In theoretical speaking, the lower the AIC, the better the regression model.

(b) Bayesian Information Criterion (BIC)

The penalty varies with the number of observations and is $\ln n$ for each parameter.

$$BIC = -2l + p \ln n ,$$

where n is the number of observations and p is the number of parameters estimated.

The table below shows the AIC and BIC value from the full and reduced model:

Model	Regressors included	AIC	BIC
Full Model	All variable	490.5450	527.7651
First Reduced Model	Remove alcohol (x_7)	488.5490	521.6335
Second Reduced Model	Remove alcohol (x_7) and systolic blood pressure (x_1)	487.9799	516.9288
Third Reduced Model	Eliminate alcohol (x_7), systolic blood pressure (x_1)and obesity (x_6)	487.6856	512.4990

In theoretical speaking, there is no value of AIC and BIC that can be considered as "good" or "bad" because AIC and BIC are focused on comparison between regression models only. AIC will penalize models that use more parameters, where if two models are explaining the same amount variation, then the model with fewer parameters will have a lower AIC score and in a better-fit model. Besides, AIC is also used to avoid overfitting while BIC is used to prevent underfitting. AIC score is low with a model with high log-likelihoods, while lower BIC indicates lower penalty terms. In addition, though BIC is always higher than BIC, lower the value of these AIC and BIC measures, better the fit model (Akaike Information Criterion, n.d.).

Referring to the table above, the full GLM model which include all of the 8 regressors is having a highest AIC and BIC value. In order to measure the goodness of fit, we start to remove the regressor which is potentially insignificant from the model by using Stepwise Backward Elimination. As a result, we found that as we remove the insignificant regressor one by one, the AIC and BIC value also decreasing eventually. After removal of all of the potential insignificant regressors (alcohol, systolic blood pressure and obesity variable) out from the model, the AIC and BIC value has improved, where it become smaller than the initial full model.

In conclusion, the third reduced model (contains 5 regressors which is significant) is determined as the best fit model among the 4 types of models mentioned above by using the penalty loglikelihood test (AIC and BIC).

4.1.2 Likelihood Ratio Test (LRT)

Likelihood Ratio Test (LRT) is a corresponding test for a generalized linear model (GLM) in logistic regression. It compares the goodness of fit of two nested regression model based on the ratio of their likelihoods, specially one obtained by maximization over the entire parameter space, and another obtained after imposing some constraints (Finnstats, 2021).

The LRT statistic is an approximate chi-square distribution with q degrees of freedom, where q is the number of constraints.

$$LRT = 2\left(\hat{l} - \tilde{l}\right)$$

Experiment 1 :

Note that the hypothesis are stated as:

H_0 : Both the full and nested models fit the data equally well. As a result, we should employ the nested model.

H_1 : The full model significantly outperforms the nested model in terms of data. As a result, We should use the entire model.

If the p-value in the test is less than a threshold of significance (0.05), then we should reject the null hypothesis (H_0) and conclude that the full model provides a significantly better fit.

Type of Model	**Regressors included**
Model 1 (Nested Model)	Excluded the alcohol variable (x_7)
Model 2 (Full Model)	All the 8 regressors are included

By conducting the Likelihood Ratio Test of Model 1 and Model 2, the result is obtained as below :

```
Likelihood ratio test

Model 1: chd ~ (sbp + tobacco + ldl + famhist + typea + obesity + alcohol +
    age) - alcohol
Model 2: chd ~ sbp + tobacco + ldl + famhist + typea + obesity + alcohol +
    age
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   8 -236.27
2   9 -236.27  1 0.0039     0.9501
```

From the output result above, we can notice that the Chi-Squared test-statistic is 0.0039 and the corresponding p-value is 0.9501. Since the p-value is more than 0.05, hence we will no evidence to reject (or will accept) the null hypothesis (H_0). This means that the full model and nested model fit the data equally well. Thus, we should use the nested model because the additional regressor variable does not offer any significant effect on improvement in the model.

Experiment 2 :

Since, the nested model (Model 1) has been chosen, then we proceed to second experiment.

Type of Model	Regressors included
Model 1 (Nested Model)	Excluded the alcohol variable (x_7)
Model 3(Nested Model)	Excluded the alcohol (x_7) and sbp (x_1) variable

By conducting the Likelihood Ratio Test of Model 1 and Model 3, the result is obtained as below :

```
Likelihood ratio test

Model 1: chd ~ (sbp + tobacco + ldl + famhist + typea + obesity + alcohol +
    age) - alcohol
Model 2: chd ~ (sbp + tobacco + ldl + famhist + typea + obesity + alcohol +
    age) - alcohol - sbp
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   8 -236.27
2   7 -236.99 -1 1.4309     0.2316
```

The Chi-Squared test-statistic is 1.4309 and the corresponding p-value is 0.2316, as shown in the output. Since the p-value of 0.2316 is greater than 0.05, hence we should choose the more "nested" model which is Model 3.

Experiment 3 :

Next, we use Model 3 to compare with another more "nested" model.

Type of Model	Regressors included
Model 3 (Nested Model)	Excluded the alcohol (x_7)and sbp variable
Model 4 (Nested Model)	Excluded the alcohol (x_7) and sbp (x_1) and obesity (x_6) variable

By conducting the Likelihood Ratio Test of Model 1 and Model 3, the result is obtained as below :

```
Likelihood ratio test

Model 1: chd ~ (sbp + tobacco + ldl + famhist + typea + obesity + alcohol +
    age) - alcohol - sbp
Model 2: chd ~ (sbp + tobacco + ldl + famhist + typea + obesity + alcohol +
    age) - alcohol - sbp - obesity
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   7 -236.99
2   6 -237.84 -1 1.7057     0.1915
```

From the output above, we can see that the Chi-Squared test-statistic is 1.7057 and corresponding p-value is 0.1915. Due to the p-value, which is less than 0.05, hence we should take model which is more "nested" (Model 4) so that it can provide a significantly better fit.

5.0 Outliers, Influential Points & Leverage Points

5.0.1 Introduction & Explanation

We will focus now on detecting potential observations that have a significant impact on the model. There are several reasons why we need to detect outliers, influential and leverage observations. First, these might be data entry errors. Secondly, influential observations may be of interest by themselves for us to study. Also, influential data points may yield biased regression coefficient estimates while leverage points may affect the model summary statistics through higher deviance values. Hence, we will identify these points using several methods.

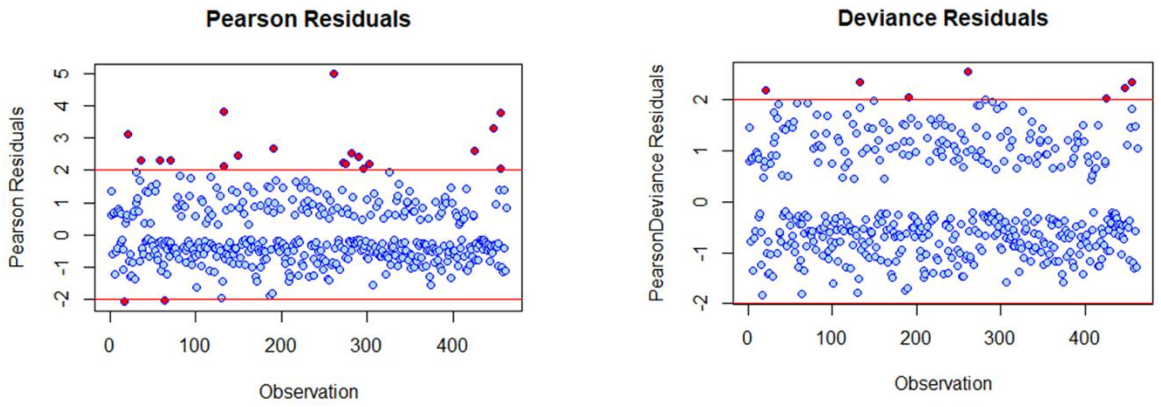
Pearson residuals are defined to be the standardized difference between the observed frequency and the predicted frequency. They measure the relative deviations between the observed and fitted values. Meanwhile, Deviance residual is another type of residual measure. It measures the disagreement between the maxima of the observed values and the fitted log-likelihood functions. It should be noted that logistic regression that uses maximum likelihood estimate (MLE) has a goal similar to the ordinary least square method (OLS), which aims to minimize the sum of the deviance residuals. Additionally, Pregibon leverage represents another key statistic that measures the leverage of an observation. These three statistics make up the three basic building blocks for logistic regression diagnostics. Thus, we will be applying the following.

5.1 Testing for Outliers

5.1.1 Pearson and Deviance Residuals

Pearson residuals are defined to be the standardized difference between the observed frequency and the predicted frequency. They measure the relative deviations between the observed and fitted values. Meanwhile, Deviance residual is another type of residual measure. It measures the disagreement between the maxima of the observed values and the fitted log-likelihood functions. It should be noted that logistic regression that uses maximum likelihood estimate (MLE) has a goal similar to the ordinary least square method (OLS), which aims to minimize the sum of the deviance residuals. We will be applying the cut-off values as shown below.

Measure	Value
leverage (hat value)	>2 or 3 times of the average of leverage
abs(Pearson Residuals)	> 2
abs(Deviance Residuals)	> 2



Based on the residual plots above, we can observe that most of the deviance and Pearson residuals for individuals with coronary heart disease (CHD) are distributed within the range of 0 to 2 whereas for individuals without CHD, the deviance and Pearson residuals are mostly distributed within the range of 0 to -2. Since most of the residuals fall within the range of 2 to -2, we can conclude that for residuals that fall out of the range of 2 to -2, they are considered as outliers and are required to be removed.

5.1.2 Remove Outliers

Refer to the previous observation, we will need to remove the outliers that will cause an impact to our model.

Based on the Pearson's residuals plot, 21 observations that are shown in the output need to be eliminated to produce a more effective model.

```
17  21  36  58  63  70 132 133 149 191 261 272 275 281 290 296 302 425 447 455 456
17  21  36  58  63  70 132 133 149 191 261 272 275 281 290 296 302 425 447 455 456
```

Whereas based on the deviance residuals plot, 7 observations that displayed below need to be eliminated.

```
21 133 191 261 425 447 456
21 133 191 261 425 447 456
```

5.2 Testing for Influential points

5.2.1 Cook's Distance

Cook's Distance is a method that used in regression analysis to find influential outliers in a set of predictor variables. It is a measure of the effect of deleting an observation on the estimated coefficients.

$$D_i(\mathbf{X}^{\mathrm{T}}\mathbf{X}, p\,\mathrm{MS}_{\mathrm{Res}}) \equiv D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)^{\mathrm{T}} \mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{p\,\mathrm{MS}_{\mathrm{Res}}}, \quad i = 1, 2, \ldots, n$$

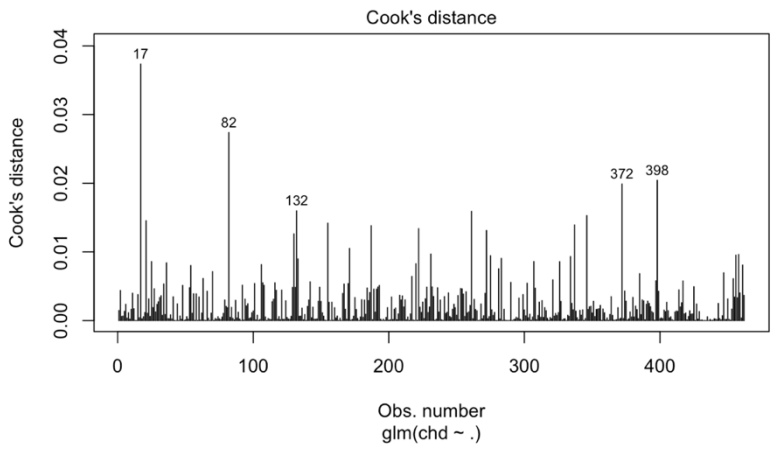
A point with larger values of D_i have considerable influence on the least-squares estimate.

Using F-distribution to interpret D_i :

A percentile of 50 indicates a highly influential point.

$F_{\alpha, p, n-p} = F_{0.5,\, 9,\, 462-9} = 0.9283567$, where α is fixed at 0.5

If $D_i > F_{\alpha, p, n-p} \rightarrow D_i > 0.9284$, then the observation i is considered as an influential point.



Based on the Cook's Distance graph above, it is clearly indicated that the maximum value of D_i value among the 462 observations is only 0.03735666, which is obviously smaller than the $F_{0.5,\,9,\,453}$= 0.9284. Hence, there is no influential point found in this data set by using method of Cook's Distance.

5.2.2 DFFITS

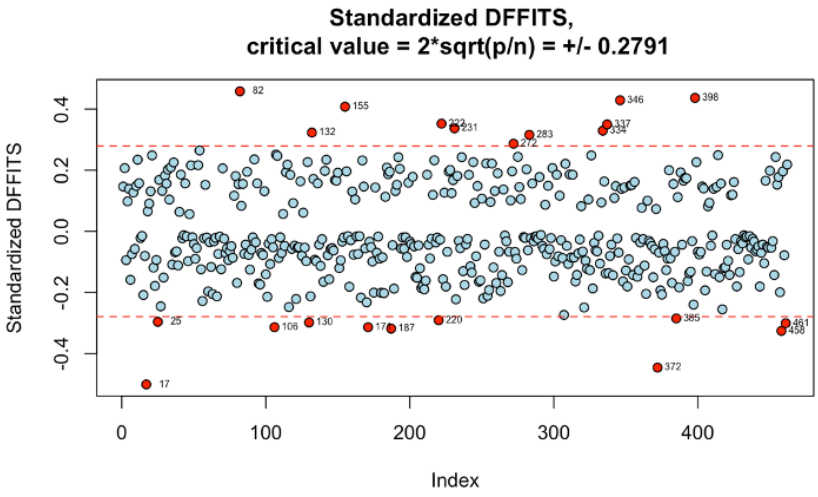
DFFITS is named as the difference in fit(s). It is considered as a studentized DDFIT. $DFFITS_i$ is the number of standard deviations that the fitted value changes if observation i is removed.

In order to determine the influence point through method of DFFITS:

Cut-off value / Threshold value $= 2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{9}{462}} = 0.2791453$

Any observation will be considered as an influential point if:

$$|DFFITS_i| > 2\sqrt{\frac{p}{n}} \rightarrow |DFFITS_i| > 0.2791453$$



Based on the observation of the graph shown above, a total number of 22 observations (red plot) are determined as a potential influential data point. They are located beyond the area bounded by two red dashed lines (cut-off value) in the graph, where the DFFITS of observations i is greater than the cut-off value. Hence, these potential influential points are needed further investigation to check whether they are significantly altering the outcome of the regression analysis.

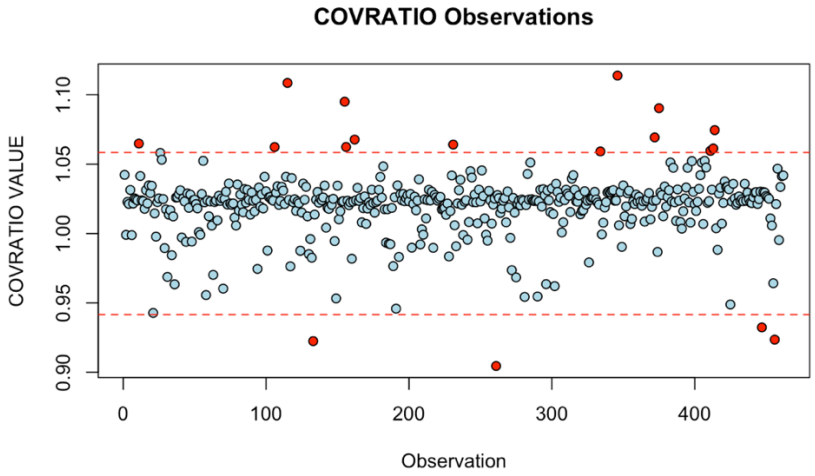
5.2.3 COVRATIO

COVRATIO is the ratio of the determinant of the coefficient covariance matrix with observation i deleted to the determinant of coefficient covariance matrix for the full model.

$$COVRATIO_i = \frac{\left|\left(\mathbf{X}_{(i)}^{\mathrm{T}}\mathbf{X}_{(i)}\right)^{-1} S_{(i)}^2\right|}{\left|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} MS_{\mathrm{Res}}\right|},\ i = 1,2,\ldots,n \qquad COVRATIO_i = \frac{\left(S_{(i)}^2\right)^p}{MS_{\mathrm{Res}}^p}\left(\frac{1}{1-h_{ii}}\right)$$

The cut-off values are $COVRATIO_i > 1 + \frac{3p}{n}$ and $COVRATIO_i < 1 - \frac{3p}{n}$.

The observation i should be considered as influential point if it meets the cut-off value of COVRATIO.



According to the graph above, we notice that the COVRATIO value of each observation is considerably evenly distributed between both cut-off value (0.9415594 and 1.058442), except a few of observations are located beyond the range. Theoretically, an observation is considered as a potential influence point if the COVRATIO value is falling either beyond 1.058442 or less than 0.9415594, which is sketched in red straight line as a cut-off point in the graph. Hence, there are influential points (red plot) among the data set which determined by COVRATIO method.

5.2.4 Remove Influential Point

Based on the observation and analysation from the **Cook's Distance graph** above, there is no influential point has been observed from the data set.

```
named integer(0)
```

On the contrary, the **method of DFFITS** has determined a total of 22 potential influential points from the data set. The following numbers are the i-th observation that are required to be removed among the 462 observations in data set:

```
17  25  82 106 130 132 155 171 187 220 222 231 272 283 334 337 346 372 385 398 458 461
17  25  82 106 130 132 155 171 187 220 222 231 272 283 334 337 346 372 385 398 458 461
```

In addition, there are 18 potential influential points discovered from the data set by using **COVRATIO method**. The following numbers are the i-th observation that are required to be eliminated among the 462 observations in data set:

```
11 106 115 133 155 156 162 231 261 334 346 372 375 411 413 414 447 456
11 106 115 133 155 156 162 231 261 334 346 372 375 411 413 414 447 456
```

5.3 Testing for Leverage Points

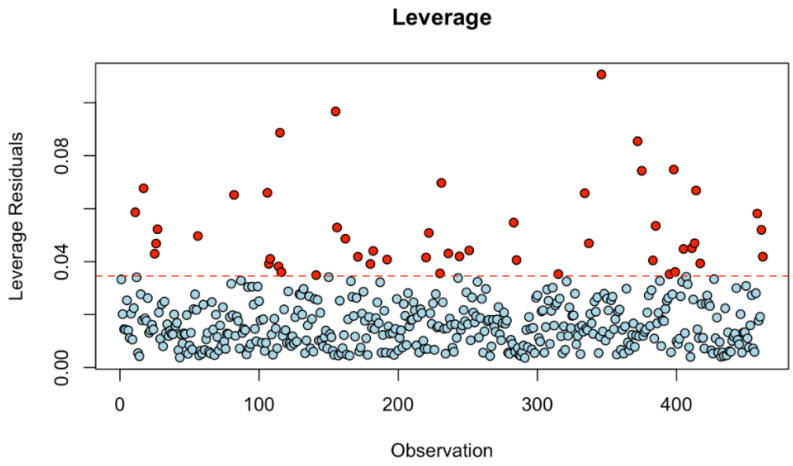
The hat matrix $H = X\left(X^{(T)}X\right)^{-1}X^T$ provides a measure of leverage. It is beneficial for examining whether certain observations deviate significantly in their X values, potentially making them exert excessive influence on the regression outcomes (Hat Matrix and Leverage - MATLAB & Simulink, n.d.).

Average hat diagonal = $\frac{p}{n}$

For any data point that has a hat diagonal value that exceeds twice the average $2\frac{p}{n}$, it is remote enough from the rest of the data to be labelled as a leverage point.

When $2\frac{p}{n} > 1$, the cut off value does not apply.

5.3.1 Remove Leverage Point



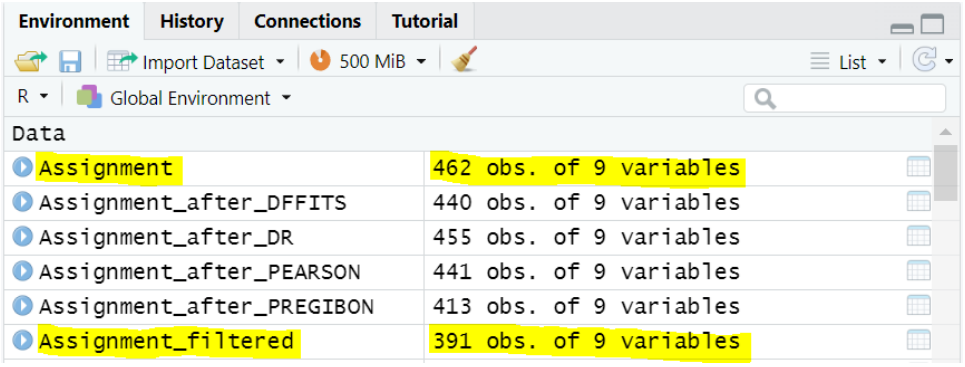
Since the maximum leverage of the model is 0.1106625 which is smaller than 1, the cut-off value $2\frac{p}{n}$ is applicable. By using the formula $2\frac{p}{n}$, we obtain a cut off value of 0.03463203. Any leverage residuals that are distributed beyond the cut-off value will be eliminated.

Based on the output, 49 observations need to be eliminated.

```
 11  17  25  26  27  56  82 106 107 108 114 115 116 141 155 156 162 171 180 182 192 220 222 230 231 236 244 251 283 285 315 334 337 346
 11  17  25  26  27  56  82 106 107 108 114 115 116 141 155 156 162 171 180 182 192 220 222 230 231 236 244 251 283 285 315 334 337 346
372 375 383 385 395 398 399 405 411 413 414 417 458 461 462
372 375 383 385 395 398 399 405 411 413 414 417 458 461 462
```

6.0 Improvement of Model based on Adjusted Dataset

6.0.1 Adjusted Dataset



After adjusting datasets for outliers, influential points and leverage points, we have removed a total of 73 observations from our initial data set. Hence, our initial dataset has reduced from 462 observations to 391 observations. However, it is important to note that the removal of observations is not an ethical way to proceed with the analysis. This is because some outliers may possess unique meanings that can only be identified if further investigation is carried out. Nonetheless, for simplicity's sake, we will assume that all anomalies identified are data entry errors and are okay to be removed.

6.1 Test: Efron's Pseudo R-squared

Efron's Pseudo R-squared is a measure of goodness-of-fit for logistic regression models. It is an alternative to the conventional R-squared used in linear regression, which cannot be directly applied to logistic regression due to the different nature of the response variable. As a reminder, our logistic regression has binary response variable which consists of 0 and 1.

Similarly, Efron's Pseudo R-squared provides an indication of the proportional reduction in the likelihood of the null model (0 regressors var. used) compared to the full model (all regressors var. used). Additionally, Efron's R-squared can be interpreted as the proportion of variability in the response variable that is explained by the predictors in the model. Thus, the higher its value, the better the model.

INITIAL DATA SET

```
$Models

Model: "glm, chd ~ ., binomial(link = \"logit\"), Assignment"
Null:  "glm, chd ~ 1, binomial(link = \"logit\"), Assignment"

$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                             0.207283
Cox and Snell (ML)                   0.234674
Nagelkerke (Cragg and Uhler)         0.323775

$Likelihood.ratio.test
 Df.diff LogLik.diff  Chisq    p.value
      -8     -61.782 123.56 6.0846e-23

$Number.of.observations

Model: 462
Null:  462

$Messages
[1] "Note: For models fit with REML, these statistics are based on refitting with ML"

$Warnings
[1] "None"

EfronRSquared
        0.245
```

ADJUSTED DATA SET

```
$Models

Model: "glm, chd ~ . - alcohol, binomial(link = \"logit\"), Assignment_filtered"
Null:  "glm, chd ~ 1, binomial(link = \"logit\"), Assignment_filtered"

$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                             0.360943
Cox and Snell (ML)                   0.351028
Nagelkerke (Cragg and Uhler)         0.502786

$Likelihood.ratio.test
 Df.diff LogLik.diff  Chisq    p.value
      -7     -84.527 169.05 3.9721e-33

$Number.of.observations

Model: 391
Null:  391

$Messages
[1] "Note: For models fit with REML, these statistics are based on refitting with ML"

$Warnings
[1] "None"

EfronRSquared
        0.372
```

The initial data set underwent evaluation using four pseudo-R-squared measures to assess the goodness of fit for a statistical model. McFadden's R-squared ($R^2_{McFadden}$) scored 0.207283, Cox and Snell's R-squared ($R^2_{Cox\ and\ Shell}$) yielded a slightly higher value of 0.234674, Nagelkerke's R-squared ($R^2_{Nagel\,\mathrm{ker}\,ke}$) displayed the highest value at 0.323772, and Efron's R-squared (R^2_{Efron}) was 0.245, indicating the model's ability to predict new data points accurately.

After applying adjustments to the data set, the pseudo-R-squared measures were revaluated to gauge the model's improved fit. The adjusted data set demonstrated substantial enhancements across all four measures. McFadden's R-squared ($R^2_{McFadden}$) increased to 0.360943, Cox and

Snell's R-squared ($R^2_{Cox\ and\ Shell}$) reached 0.351028, and Nagelkerke's R-squared ($R^2_{Nagel\,\mathrm{ker}\,ke}$) rose to 0.502756. Furthermore, Efron's R-squared (R^2_{Efron}) increased to 0.372, indicating the model's enhanced predictive accuracy. These augmented pseudo-R-squared values reveal the improved explanatory power and predictive capabilities of the adjusted model, providing valuable insights for researchers and analysts seeking robust inferences from the data.

6.2 Test: ROC and AUC Curve

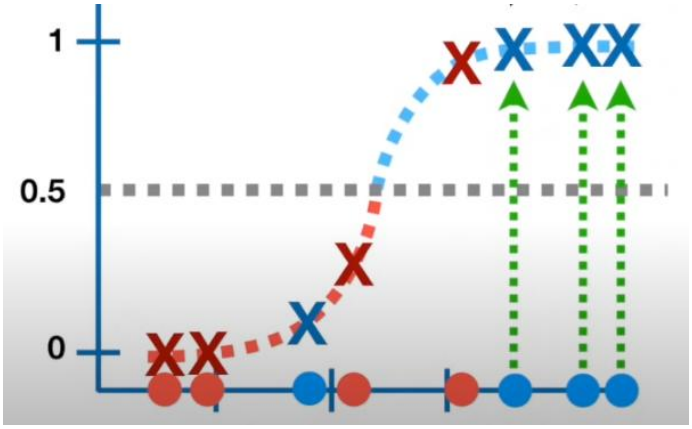
ROC stands for "Receiver Operating Characteristic" while AUC stands for "Area Under the Curve". We have chosen this method as it is a commonly used evaluation metrics for binary classification models. Basically, the ROC curves visually assess the model's performance while AUC provides a single scalar value to quantify the model's ability to separate classes. The ROC curve simply makes up the probability curve while AUC is just the area under this ROC curve. Typically, the large the AUC, the better the model. One way to look at it is that an area of 1 is a perfect model where there exists a point at coordinates (0,1). At (0,1), it would indicate that FPR = 0 and TPR = 1, meaning that all positive values are predicted correctly while no negative values are predicted wrongly.

The ROC curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds for a binary classification model. Let's have this explained visually below.

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

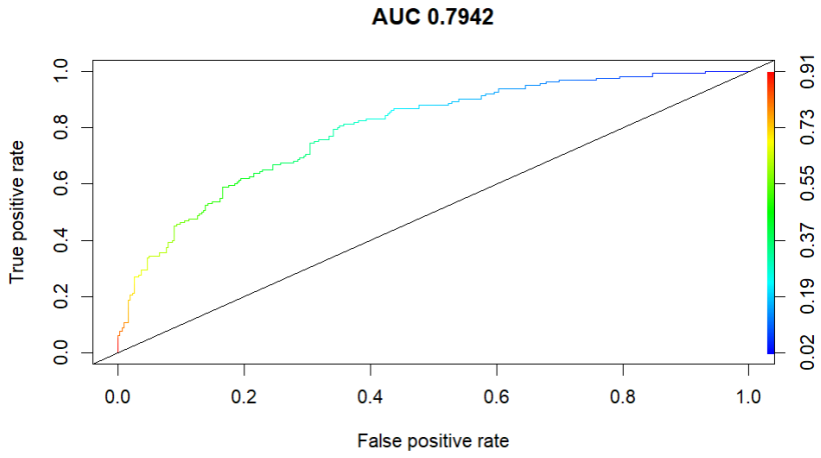
We obtain the following formulas:

$$TPR = \frac{TP}{TP} + FN \qquad FPR = \frac{FP}{FP} + TN$$



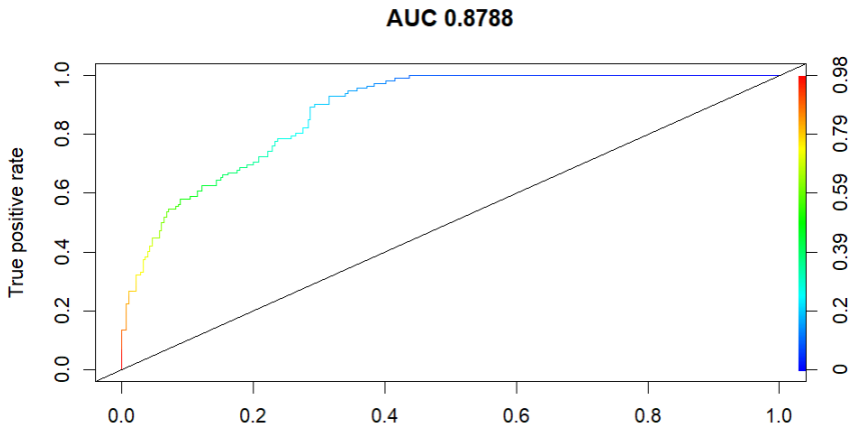
Through binomial canonical link formula, we would produce a logit mean function that looks like above S-curve. Additionally, the grey horizontal line represents the threshold line. For each observation, they will have a corresponding predicted probability which is the y-value of this curve. Hence, for our dataset, a predicted probability of 0.5 and above would indicate that the observations would have CHD while 0.5 and below would indicate no CHD. From here, TPR and NPR can be calculated. Following this, the grey threshold line can take a value from 0 to 1, where anything below the threshold line represents no CHD while anything above will represent have CHD. Thus, this will produce many different combinations of TPR and NPR. When plotted on the same graph and connected in a line, it will produce the ROC curve.

ROC AUC Curve (INITIAL DATA SET)



Notice that the initial data set with a full model has an AUC value of 0.7942.

ROC AUC Curve (ADJUSTED DATA SET)



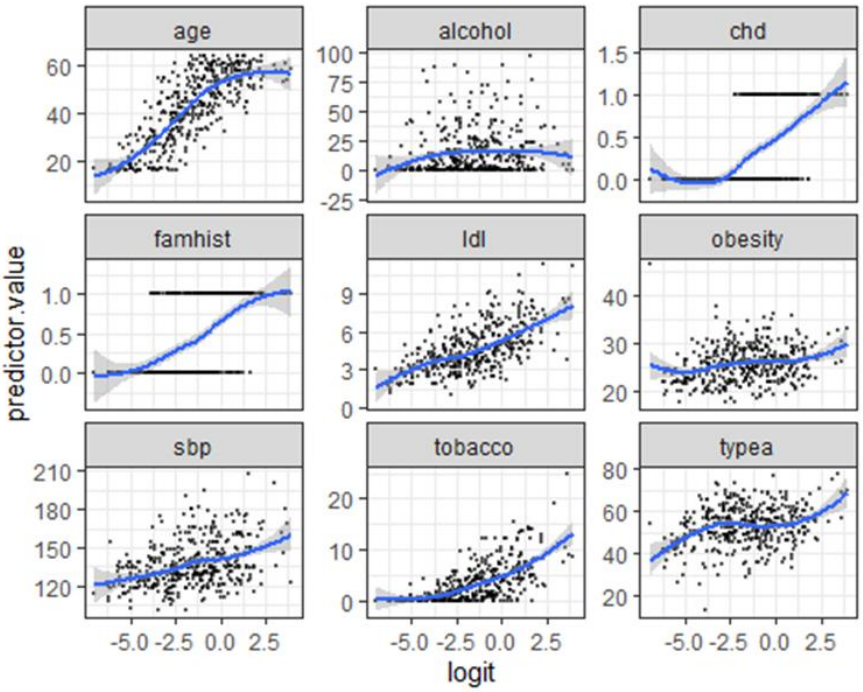
Notice that the adjusted data set shows a significant improvement in the AUC value. Hence, it is concluded that the adjusted data set is a better model. Additionally, the adjusted dataset model shows a better discriminate power where it is proficient at distinguishing between positive cases and negative cases. In Layman's terms, the model is effective in accurate predictions.

7.0 GLM for Adjusted Data

7.1 Assumptions for Binomial Logistics Regression

Through our adjusted data, we have complied to all six assumptions that are required before using a binomial logistics regression. Initially, for our dataset, we ignored anomalies (outliers, influential and leverage points); thus, violating the 6th assumption. However, through our adjusted dataset, we have adhered to this assumption.

As we have already verified the first 3 assumptions required for binomial logistics regression for the initial dataset, these 3 assumptions should remain valid since we are just removing observations and not regressor variables.



After running the same logistics regression diagnostics, we observe that a linear relationship exists between each continuous independent variable and the logit transformation of the dependent variable. Thus, 4th assumption is complied. The 5th assumption regarding the absence of multicollinearity is also respected. In Content 7.5, we have calculated the Variance Inflation Factor (VIF) where all the values are much lower than 5.

Thus, our adjusted dataset proves to respect all six assumptions, deeming binomial logistic regression a highly compatible method with our dataset. Hence, results of high quality, credibility and validity will be produced.

7.2 Full GLM Model

The summary statistics from the full fitted GLM model will be shown as below:

```
Call:
glm(formula = chd ~ ., family = binomial(link = "logit"), data = Assignment_filtered)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.00e+01   1.97e+00   -5.09  3.6e-07 ***
sbp          1.13e-02   8.74e-03    1.30  0.19484
tobacco      1.41e-01   4.11e-02    3.42  0.00063 ***
ldl          3.81e-01   9.14e-02    4.17  3.1e-05 ***
famhist      1.40e+00   2.97e-01    4.70  2.6e-06 ***
typea        6.21e-02   1.81e-02    3.44  0.00059 ***
obesity     -8.43e-02   4.38e-02   -1.93  0.05421 .
alcohol     -6.43e-04   7.82e-03   -0.08  0.93453
age          7.05e-02   1.47e-02    4.79  1.6e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 468.37  on 390  degrees of freedom
Residual deviance: 299.31  on 382  degrees of freedom
AIC: 317.3

Number of Fisher Scoring iterations: 6
```

The adjusted fitted model equation is stated as below:

$$y = -10.01 + 0.0113x_1 + 0.141x_2 + 0.381x_3 + 1.40x_4 + 0.0621x_5 - 0.0843x_6 - 0.000643x_7 + 0.0705x_8$$

Coefficients	Estimate	Standard Error	Z value	Pr(>|z|), p-value
Intercept	-10.01	1.97	-5.09	0.000000356
Systolic Blood Pressure (SBP)	0.0113	0.00874	1.30	0.195
Tobacco	0.141	0.0411	3.42	0.000633
Low-density lipoproteins cholesterol (LDL)	0.381	0.0914	4.17	0.0000309
Family History	1.40	0.297	4.70	0.00000257
Type A Personality	0.0621	0.0181	3.44	0.000591
Obesity	-0.0843	0.0438	-1.93	0.0542
Alcohol	-0.000643	0.00782	-0.0821	0.935
Age	0.0705	0.0147	4.79	0.00000163

7.3 Reduced GLM (Adjusted dataset)

Using stepwise regression formula, we obtained the following summary statistics. Note that the result shown below is the reduced GLM which is excludes Systolic Blood Pressure (x_1) and Alcohol variable (x_7).

```
Call:
glm(formula = chd ~ tobacco + ldl + famhist + typea + obesity +
    age, family = binomial(link = "logit"), data = Assignment_filtered)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.6625     1.6409   -5.28  1.3e-07 ***
tobacco       0.1408     0.0396    3.56  0.00038 ***
ldl           0.3816     0.0901    4.23  2.3e-05 ***
famhist       1.3793     0.2951    4.67  2.9e-06 ***
typea         0.0590     0.0178    3.31  0.00092 ***
obesity      -0.0773     0.0430   -1.80  0.07221 .
age           0.0749     0.0143    5.25  1.5e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 468.37  on 390  degrees of freedom
Residual deviance: 301.04  on 384  degrees of freedom
AIC: 315

Number of Fisher Scoring iterations: 6
```

7.4 Measure of Fit (Adjusted dataset)

7.4.1 Test: AIC & BIC

The table below shows the AIC and BIC value from the full and reduced model:

Model	Regressors included	AIC	BIC
Full Model	All variable	317.31	353.03
First Reduced Model	Remove alcohol (x_7) variable	315.31	347.06
Second Reduced Model	Remove alcohol (x_7) and systolic blood pressure (x_1) variable	315.04	342.82

Note that the AIC and BIC values continue to drop as we remove the regressors based on the similar procedure of stepwise regression methods. It should be also noted that the AIC and BIC value of adjusted dataset is much lower than initial dataset. For example, full model of AIC for initial dataset and adjusted dataset is 490.5450 and 317.31 respectively. Nonetheless, there is no particular meaning to AIC for comparison between different data sets. Yes, the AIC value

can change for increased number of observations. However, AIC is self-referential, which means that one can only compare different models using the SAME data set, not different data sets.

7.5 Test for Multicollinearity: VIF (Adjusted dataset)

R code to conduct VIF Test for adjusted full GLM model:

```
###
car::vif(fullModel)
```

Result obtained from VIF Test:

```
     sbp tobacco     ldl famhist   typea obesity alcohol     age
  1.1774  1.1161  1.1870  1.0296  1.1744  1.2092  1.1388  1.2810
```

The VIF values for each regressor variable has a maximum value of 1.2810. These values are far below 5, which is the value that indicates a concern for multicollinearity. Hence, we can conclude that our adjusted dataset has minimal multicollinearity feature.

8.0 Conclusion

In the initial data, we have identified that there are certain variables, namely x1, x6, and x7, needed to be removed from the dataset. After adjusting our dataset, we observed that removing x1 and x7 sufficed to enhance the model's accuracy and interpretability. Retaining the relevant information from x6 in the adjusted dataset allowed us to achieve a more refined model that better captured the relationships between the remaining predictors and the response variable.

Initial Dataset Fitted Equation:

$$y = -6.4169865 + 0.0795655x_2 + 0.1824114x_3 + 0.9234083x_4 + 0.0389489x_5 + 0.048956x_8$$

Adjusted Dataset Fitted Equation:

$$y = -10.01 + 0.141x_2 + 0.381x_3 + 1.4x_4 + 0.0621x_5 - 0.0843x_6 + 0.0705x_8$$

Nevertheless, this report acknowledges several limitations. Firstly, the dataset might lack representation of certain demographic groups, which could limit the generalizability of the findings. If the data predominantly focuses on a specific population, region or ethnicity, it may not fully capture the diverse factors influencing CHD across various demographics. This limitation warrants caution when interpreting the results and making broad inferences about the entire population.

Besides that, If Body Mass Index (BMI) is used as a metric, it may not provide a complete picture of an individual's health status. For example, individuals with a high BMI might have elevated mass due to increased muscle mass, particularly in physically active individuals, rather than excess body fat. In such cases, relying solely on BMI to assess obesity could be misleading, as it may misclassify individuals who are healthy and physically active. To avoid potential misinterpretations, we should be mindful of the limitations of BMI and consider additional measures that capture variations in body composition accurately.

Other than that, the dataset of the alcohol regressor variable exhibits an uneven range distribution, indicating variations in the spread and magnitude of the data points. To illustrate, 70% of the alcohol data corresponds to the first 15% of the sampled range. The uneven distribution of the regressors' values can have implications for statistical analyses and modelling. It may lead to disparities in the regressor coefficients, which will lead to a potentially biased response variable and model's overall performance.

Another noteworthy limitation of the CHD dataset is the potential presence of missing or incomplete data. Missing data can introduce bias and compromise the validity of the analysis if not appropriately addressed. The impact of missing data on the results needs careful consideration, and researchers may need to implement suitable imputation techniques or handle incomplete cases judiciously to ensure the robustness of the conclusions. Rigorous methods for handling missing data are essential to avoid skewed or misleading interpretations of the relationships between predictors and CHD outcomes.

9.0 Reference

Centers for Disease Control and Prevention. (2021, July 19). *Coronary Artery Disease (CAD*). https://www.cdc.gov/heartdisease/coronary_ad.htm#:~:text=Coronary%20artery%20disease%20is%20caused,This%20process%20is%20called%20atherosclerosis.

Centers for Disease Control and Prevention. (2023, May 3). *Family health history of heart disease.* https://www.cdc.gov/genomics/disease/fh/history_heart_disease.htm#:~:text=If%20you%20have%20a%20family,cholesterol%2C%20can%20run%20in%20families

Cross Validated. (2018, Nov 8). *How is the Akaike information criterion (AIC) affected by sample size?* https://stats.stackexchange.com/questions/376059/how-is-the-akaike-information-criterion-aic-affected-by-sample-size#:~:text=Yes%2C%20the%20AIC%20value%20can,set%2C%20not%20different%20data%20sets.

Foley, M. (2019, May 31). *How to handle influential data points.* https://rpubs.com/mpfoley73/501093

Johns Hopkins Medicine (n.d.). *Coronary Artery Disease* https://www.hopkinsmedicine.org/about/_downloads/building-healthy-communities/cad-risk-factors.pdf

Jousilahti, P. & et al. (1999, March 9). *Sex, age, cardiovascular risk factors, and coronary heart disease.* https://www.ahajournals.org/doi/10.1161/01.CIR.99.9.1165

Mai, Y. & Zhang, Z. (2017*). Statistical Power Analysis for One-way ANOVA with Binary or Count Data.* https://webpower.psychstat.org/wiki/_media/grant/mai-zhang-2017.pdf

National Center for Biotechnology Information. (n.d.). *Cardiovascular diseases - How tobacco smoke causes disease: The biology and behavioral basis for smoking-*

attributable disease - NCBI bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK53012/

OARC Stats – Statistical Consulting Web Resources – UCLA. (n.d.*). Lesson 3 logistic regression diagnostics*. https://stats.oarc.ucla.edu/stata/webbooks/logistic/chapter3/lesson-3-logistic-regression-diagnostics-2/

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

PubMed Central (PMC). (2007, September 8). *Families of patients with premature coronary heart disease: An obvious but neglected target for primary prevention.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1971158/

PubMed Central (PMC). (2017). *Risk factors for coronary artery disease: Historical perspectives.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5686931/

PubMed Central (PMC). (n.d.). *Growing epidemic of coronary heart disease in low- and middle-income countries*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2864143/

PubMed Central (PMC). (n.d.). *Obesity in coronary heart disease: An unaddressed behavioral risk factor.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5640469/

Raoniar R. (2020, March 10). *Modelling Binary Logistic Regression using R.* https://onezero.blog/modelling-binary-logistic-regression-using-r-research-oriented-modelling-and-interpretation/

Salamon, M. (2021, January 29). *How alcohol affects heart failure*. WebMD. https://www.webmd.com/heart-disease/heart-failure/features/alcohol-heart-failure#:~:text=A%20lot%20of%20research%20has

ScienceDirect. (n.d.). *Akaike Information Criterion.* https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/akaike-information-criterion

SPSS Statistics Tutorials and Statistical Guides (n.d.*). Binomial Logistic Regression using Stata.* https://statistics.laerd.com/stata-tutorials/binomial-logistic-regression-using-stata.php#:~:text=A%20binomial%20logistic%20regression%20is,referred%20to%20as%20logistic%20regression

Starmer, J. (2019). *ROC and AUC, Clearly Explained!.* https://www.youtube.com/watch?v=4jRBRDbJemM

STHDA (2019, February 1). *Logistic regression assumptions and diagnostics in R.* http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/

Turney, S. (2022, May 23). *Chi-Square (X^2) Tests | Types, Formula & Examples*. Scribbr. https://www.scribbr.com/statistics/chi-square-tests/

WebMD. (2020). Facebook.com. https://www.facebook.com/WebMD.

Zach. (2021, May 20). *What is Considered a Good AIC Value?* Statology. https://www.statology.org/what-is-a-good-aic-value/