Project 2 Report: Cardiovascular Disease Logistic Regression

Manpreet Dhindsa, Sarah Rodgers, Nitika Kataria, Amber Curran

mkd8bb, pjk2wq, nk3rf, akc6be

Executive Summary

Heart disease is one of the leading causes of death in the United States with approximately 655,000 deaths every year. With cardiovascular disease being so prevalent in our country, our group was interested in better understanding the factors that may lead to someone developing cardiovascular disease in their lifetime. In order to do so, we utilized the Kaggle Cardiovascular Disease dataset that contains various medical information and factors for 70,000 patients. This data was collected for each person during a medical examination visit with various features related to heart disease that span objective, examination, and subjective inputs. Objective inputs are factual information, examination inputs are results determined from the examination of the person, and lastly, subjective inputs are information provided by the person about themselves. Our goal is to build a logistic regression model to understand which factors have a greater influence on the presence or absence of cardiovascular disease. We concluded that age, height, weight, diastolic blood pressure, cholesterol, smoking, and being active are influential predictors while in the presence of one another for determining the presence or absence of cardiovascular disease and that this model is useful in identifying people who did and did not develop cardiovascular disease. These predictors that correspond with the common factors of people who develop heart disease through research are high blood pressure, high cholesterol, and smoking. ("Know Your Risk for Heart Disease") We also determined that no predictors are heavily correlated with one another, which contradicted our original hypothesis of systolic and diastolic blood pressure being heavily correlated.

The final model we found is:

log (odds) = -4.8936545 + 0.0901056*(Age) - 0.0099160*(Height) + 0.0094765*(Weight) + 0.0014699*(Diastolic blood pressure) + 0.6646955*(If that individual has high cholesterol) - 0.1887472*(If that individual smokes)- 0.2107440*(If that individual is active).

This helps illustrate that as an individual gets older, gains more weight, has a higher diastolic blood pressure, has high cholesterol and/or smokes, the odds of having cardiovascular disease increases. Conversely, the odds of having heart disease may decrease the taller an individual is and/or if that person is active. Therefore, we have concluded that a person can most effectively and realistically decrease their risk of heart disease by monitoring their weight, reducing their diastolic blood pressure and cholesterol, not smoking, and having an active lifestyle.

Exploratory Data Analysis

The cardiovascular disease dataset contains records for 70,000 subjects with one binary response variable, presence or absence of cardiovascular disease (Cardio) with 0 representing people who did not develop cardiovascular disease and 1 representing people who did. The following predictors listed in Tables 1 and 2 were included in this dataset:

Table 1: Cardiovascular Disease Continuous Variables

Variable Label	Variable Name	Values	Mean
Age	Age	29 to 64	52.80
Height	Height (cm)	55 to 250	164.36
Weight	Weight (kg)	10 to 200	74.21
AP_HI	Systolic blood pressure	-150 to 16,020	128.82
AP_LO	Diastolic blood pressure	-70 to 11,000	96.63

Table 2: Cardiovascular Disease Categorical Variables

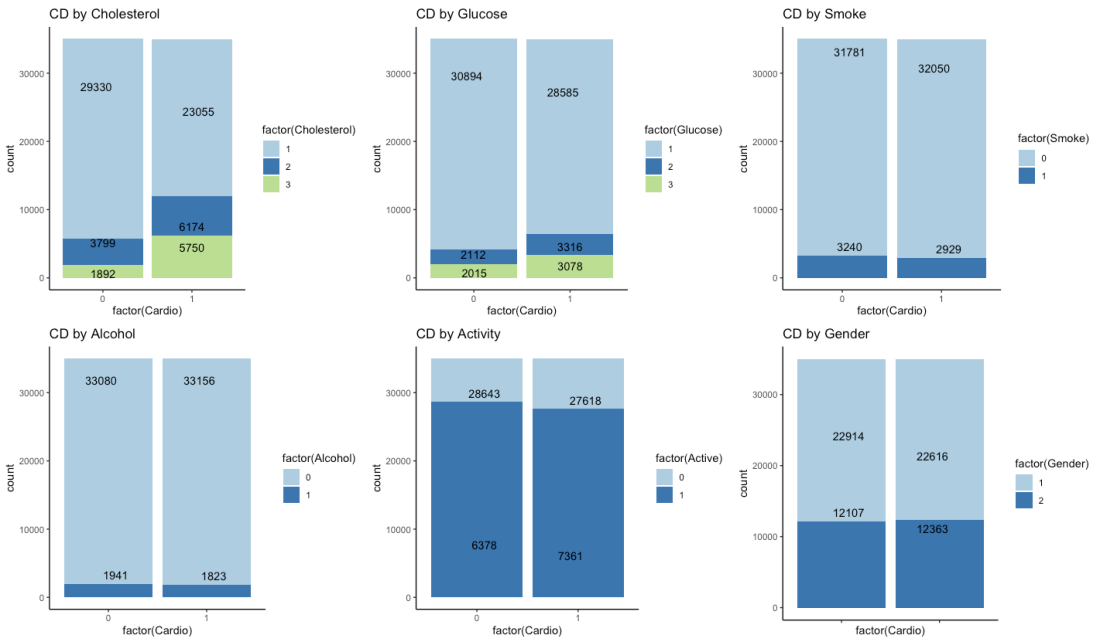
Variable Label	Variable Name	Variable Levels	Reference Class
Gender	Gender	1: Women 2: Men	Women
Cholesterol	Cholesterol	1: Normal 2: Above normal 3: Well above normal	Normal
Glucose	Glucose	1: Normal 2: Above normal 3: Well above normal	Normal
Smoke	Smoking	0: Does not smoke 1: Does smoke	Does not smoke
Alcohol	Alcohol intake	0: Does not intake alcohol 1: Does intake alcohol	Does not intake alcohol
Active	Physical activity	0: Does not perform physical activity 1: Does perform physical activity	Does not perform physical activity

Prior to performing EDA, we performed data cleansing activities to ensure the data was in expected formats and prepared properly for model building. This included storing the data of each variable into different columns since the data import into R stored the values in one single column. We also renamed and changed variables to be in the appropriate formats of numeric or factor. Additionally, we noticed each patient's age was recorded in days, so we converted age to be in years instead. Lastly, we identified that there was no missing data in the dataset.

We performed EDA to determine how the predictors relate to people who developed cardiovascular disease. We first noticed that the distribution of people who did and did not develop cardiovascular disease within the dataset are approximately equal with 34,979 and 35,021 people, respectively. After doing so, we plotted each of the categorical variables against

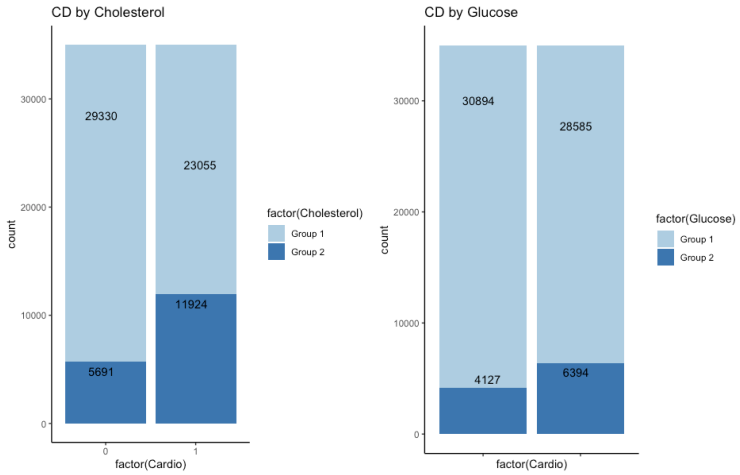
the response variable, Cardio, to understand the distributions of each value. As seen below in Figure 1, the categorical variables of smoking, alcohol, activity, and gender are approximately equal across people who did and did not develop cardiovascular disease.

Figure 1: Cardiovascular Disease Distribution by Categorical Predictors



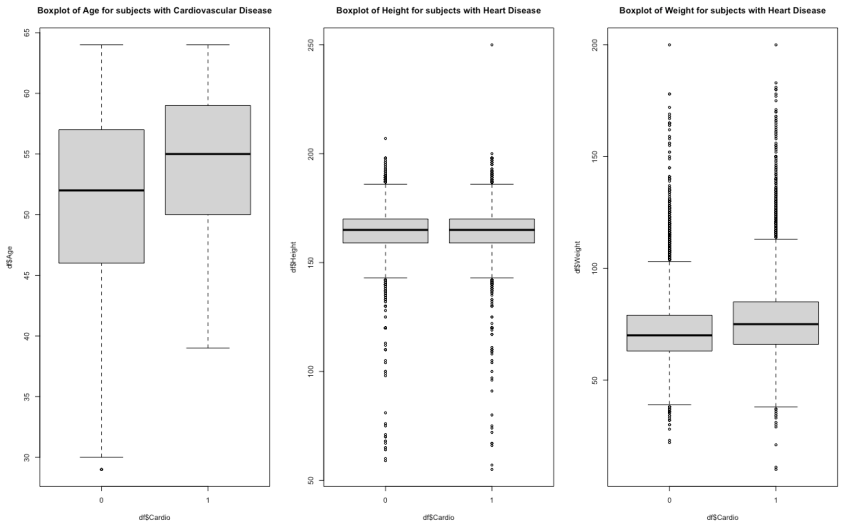
When assessing cholesterol and glucose, it can be seen that the above normal group (2) and well above normal group (3) have a smaller frequency compared to the normal group (1). Therefore, we decide to group together the above normal (2) and well above normal (3) levels in the cholesterol and glucose variables. The grouped values for cholesterol and glucose are shown below in Figure 2 and will be used in subsequent analyses. Here we can see those individuals who developed cardiovascular disease appear to have higher cholesterol and glucose values than those who did not develop cardiovascular disease.

Figure 2: Cardiovascular Disease Distribution by Cholesterol and Glucose Grouped



We also plotted each of the numerical predictors against the response variable using boxplots to understand the distributions, and identify possible outliers. It can be seen in Figure 3, that the people who did develop cardiovascular disease appear to be older and weigh more than people who did not develop cardiovascular disease.

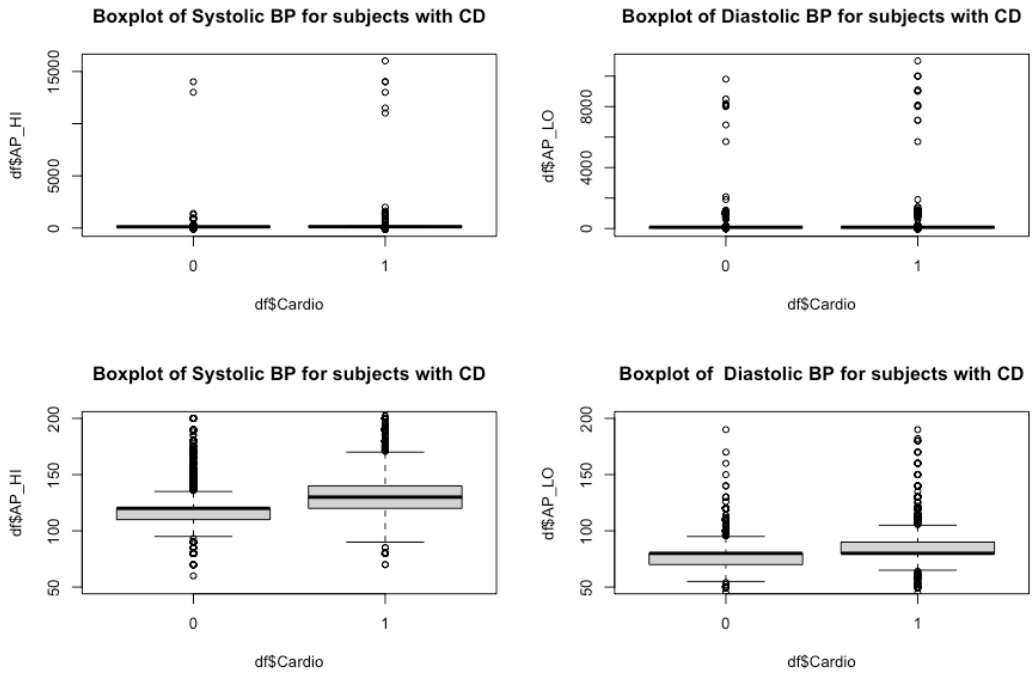
Figure 3: Cardiovascular Disease Distribution by Age, Height, and Weight



The last distribution assessed was systolic and diastolic blood pressure on the response. The first row of boxplots in Figure 4 depicts the distribution of these variables which shows the wide

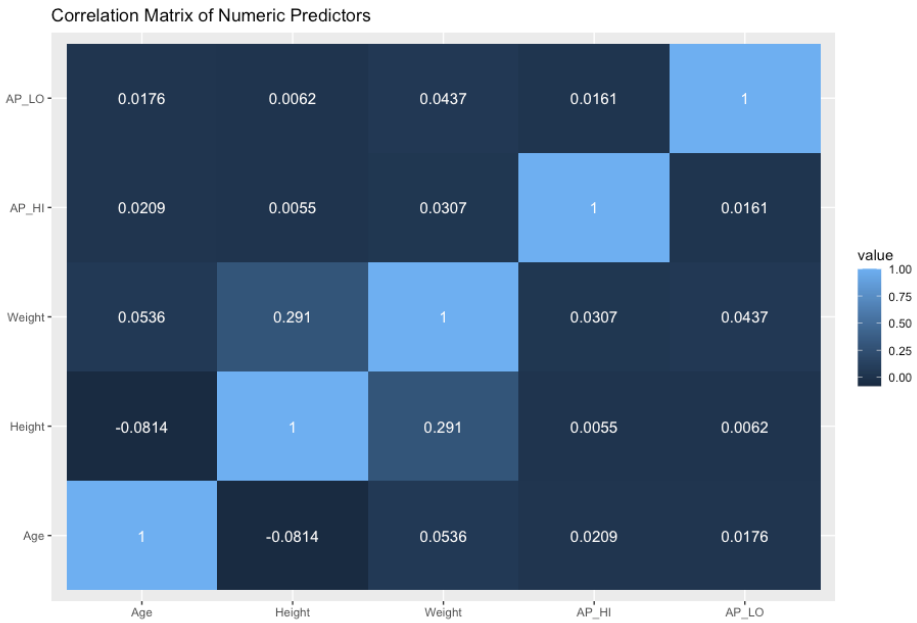
range of relevant variables and possible outliers. In order to assess the interquartile range, we created a second set of boxplots for both blood pressure variables with a specified y-axis range. Here we can see that people who did develop cardiovascular disease have higher systolic and diastolic blood pressure than people who did not develop cardiovascular disease.

Figure 4: Cardiovascular Disease Distribution by Systolic and Diastolic Blood Pressure



We then developed a correlation matrix to identify if any of the numerical predictors were heavily correlated with one another. Initially we presumed that the systolic blood pressure and diastolic blood pressure variables would be highly correlated, however as seen below in Figure 5, there are no numerical predictors that are heavily correlated with one another.

Figure 5: Correlation Matrix of the Numeric Predictors

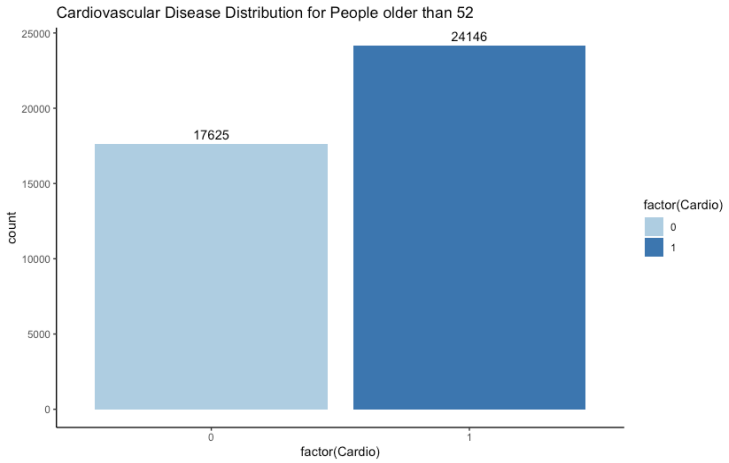


As mentioned when assessing Figure 4, there appears to be possible outlying values as there are observations with negative blood pressure values and values that are above 10,000. Prior to assessing these variables further, we wanted to understand the difference between systolic and diastolic blood pressure so that we are more confident in our interpretations. Systolic blood pressure is the maximum pressure the heart exerts while beating whereas diastolic blood pressure is the amount of pressure in the arteries between heart beats. ("Cardiovascular Disease dataset") Since these variables are very closely related, we decided to not consider or utilize the systolic blood pressure variable in modeling our data as it causes issues while building the logistic regression model. More specifically, systolic blood pressure is essentially another form of the response variable, and therefore, including it in the model will be redundant as the conclusion we are trying to understand is already provided as a predictor variable. Since blood pressure results are expected to be no higher than 180 and 120 mmHg for systolic and diastolic readings respectively, we can identify there are readings very elevated that may influence the model. We

performed Cook's Distance to determine if there are influential outliers as this test measures how removing an observation changes the predicted values for all of the observations. We determined there are no influential observations with 11 predictors and 70,000 observations.

Further research suggested that people ages 65 and older are much more likely to suffer a heart attack, have a stroke, or develop coronary heart disease or heart failure than younger people. (“Heart Health and Aging ") With this in mind and in conjunction with the EDA, we subsetted the cardiovascular dataset to remove the younger population and only include people older than 52 years old. This threshold was chosen because the mean age of the population of people in our dataset was 52 years old. After subsetting the dataset there were 41,771 observations left to utilize in the model building process. In Figure 6, it can be seen that there still is a large distribution of people 52 years or older who did and did not develop cardiovascular disease.

Figure 6: Distribution of people 52 years old and older against Cardiovascular Disease



We then performed a similar EDA as previously described on the entire population, but only on those individuals 52 years or older. When assessing the distributions provided below in Figures 7, 8 and 9, similar insights were determined for the 52 years or older population as the whole population. More specifically, people who are 52 years or older that developed cardiovascular

disease tend to have higher cholesterol, glucose, age, weight, systolic and diastolic blood pressure when compared to those that did not develop cardiovascular disease.

Figure 7: CD Distribution by Categorical Predictors for People 52 years or older

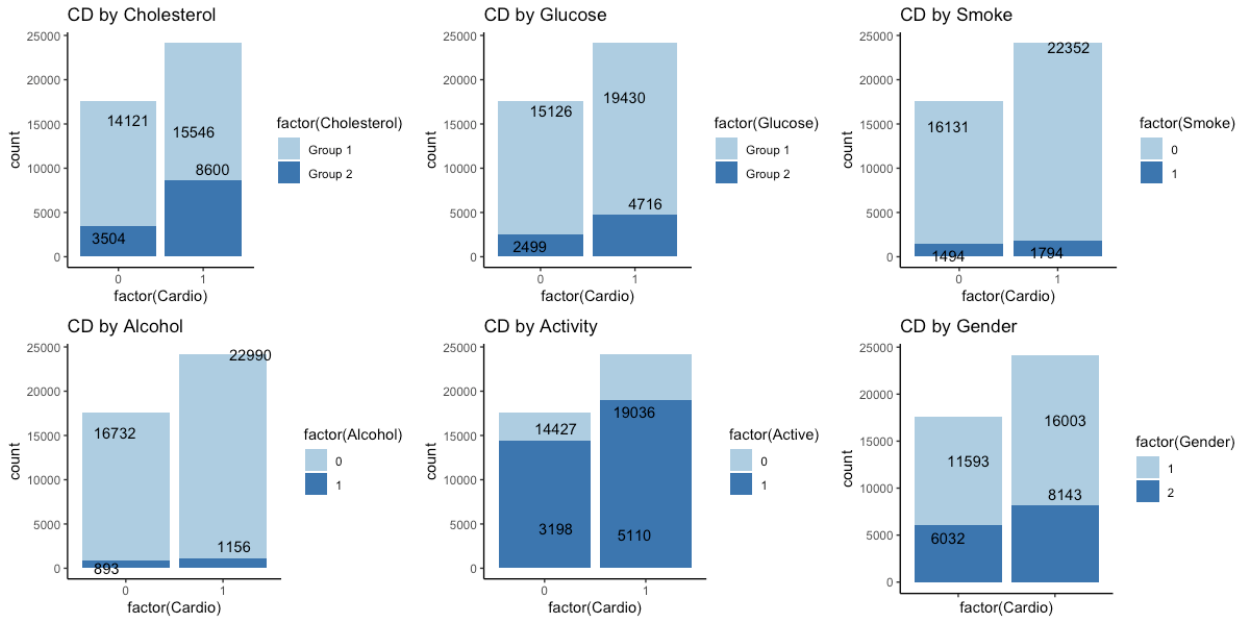


Figure 8: CD Distribution by Age, Height, and Weight for People 52 years or older

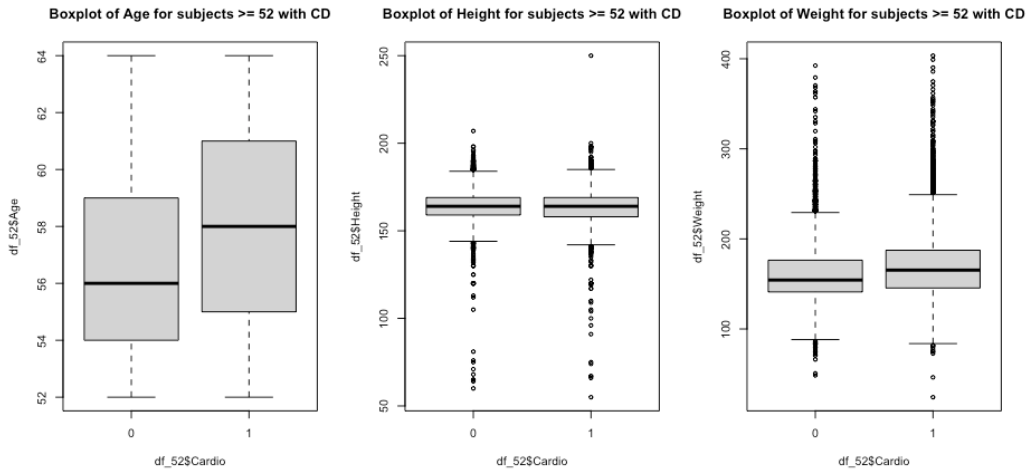
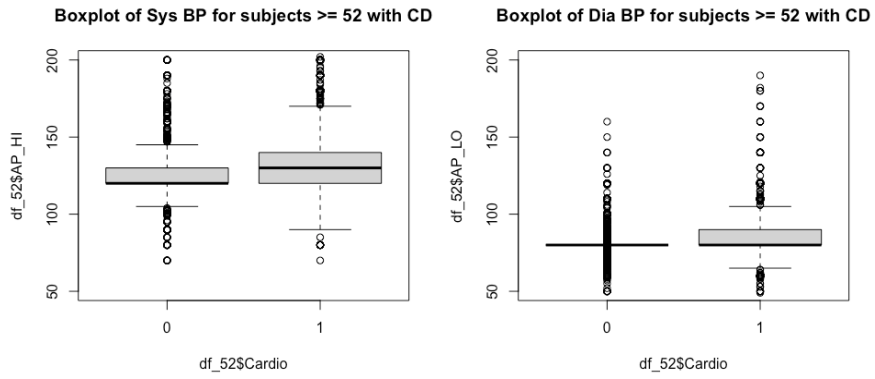


Figure 9: CD Distribution by Systolic & Diastolic Blood Pressure for People 52 years or older



Building a Model

Considering only the population of people 52 years and older, the data was split in half including a test dataset and a training dataset using the set seed 111. We first verified the distribution of people in the testing and training dataset for those with and without cardiovascular disease to ensure there were enough observations in both classes and the number of people were relatively even between the datasets.

To build the model itself we used the training dataset as this was determined to be an adequate dataset to predict the rate of cardiovascular disease. First, a model was fit performing all possible regressions including all predictors in the dataset except systolic blood pressure (AP_HI) which was determined to be dropped based on the exploratory data analysis. Then, a data frame was created to store the predictors in the various models considered as well as their various criteria. With people older than 52, the predictors that led to a first-order model having the best adjusted R^2, the MS_{RES}, and the Mallow's C_P included age, height, weight, diastolic blood pressure (AP_LO), cholesterol, smoke, alcohol, and active only. The best model according to the BIC includes the same predictors without alcohol. This suggests that the predictors gender and

glucose are not significantly contributing to the model and can be dropped as predictors which we will confirm below by conducting the appropriate hypothesis test.

Additionally we utilized the automatic search procedures of forward selection, backward elimination, and stepwise regression to determine which predictors are best to be used in the model. For forward selection, the model started with the intercept only model and continued to add predictors based on their AIC values and continued to add one predictor at a time until the AIC no longer improved with the addition of a predictor. Based on forward selection, the best model would include all predictors. For backward elimination, the model started with the full model including all the predictors and continued to remove one predictor at a time that was least significantly contributing to the model until a best model was determined. Based on backward elimination, the best model would include all predictors except glucose and gender as predictors. For stepwise regression, the model worked in adding and removing predictors simultaneously to determine which predictors were significantly contributing to the model the most to create a best model. Based on stepwise regression, the best model would include all predictors except glucose and gender as predictors, similar to what was seen for backward elimination. As you can see with each of these automated procedures, different models are determined to be the best model and it is difficult to guarantee one “best” model. With this in mind these models are a good starting point, and we determined the best model in this case includes dropping the predictors gender and glucose and to consider only predictors that are significantly contributing to the model.

Hypothesis Tests

When assessing the results from the model selection criteria and the automatic search procedures, many of the results did not lead to the same conclusions, as expected. This is a typical concern when using these model selection methods, so additional time should be taken to

determine which of the results to use. To ensure our model contains all the necessary predictors, we conducted a few hypothesis tests. The hypothesis tests consisted of the following for our data set:

- Hypothesis test utilizing Chi-Square Model Deviance to drop *Gender* and *Glucose*:

  $H_0 : \beta_9 = \beta_{10} = 0; H_a :$ *At least one predictor in the null hypothesis is not zero*

  Using the residual deviance from Figure 10 and the residual deviance from Figure 11, the chi-square model deviance can be found. The p-value, determined using the chi-square model, is **1.** Therefore, with a confidence level of *0.05*, we fail to reject the null hypothesis and conclude that *Gender* and *Glucose* are insignificant predictors while in the presence of the other predictors and can be dropped from the model.

Figure 10: Summary of Model containing Gender and Glucose

```
Call:
glm(formula = train$Cardio ~ train$Age + train$Height + train$Weight +
    train$AP_LO + train$Cholesterol + train$Smoke + train$Alcohol +
    train$Active + train$Gender + train$Glucose, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8523  -1.1625   0.7267   1.0315   1.7032

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -4.7913411  0.4254991 -11.261  < 2e-16 ***
train$Age                  0.0898993  0.0041880  21.466  < 2e-16 ***
train$Height              -0.0105364  0.0021577  -4.883 1.04e-06 ***
train$Weight               0.0210010  0.0011521  18.228  < 2e-16 ***
train$AP_LO                0.0014678  0.0001766   8.309  < 2e-16 ***
train$CholesterolGroup 2   0.6785904  0.0366189  18.531  < 2e-16 ***
train$Smoke1              -0.1721758  0.0595244  -2.893  0.00382 **
train$Alcohol1            -0.1161590  0.0705079  -1.647  0.09946 .
train$Active1             -0.2099574  0.0368464  -5.698 1.21e-08 ***
train$Gender2              0.0263389  0.0369629   0.713  0.47611
train$GlucoseGroup 2      -0.0328178  0.0432487  -0.759  0.44796
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28377  on 20884  degrees of freedom
Residual deviance: 26750  on 20874  degrees of freedom
AIC: 26772
```

Figure 11: Summary of Model without Gender and Glucose

```
Call:
glm(formula = train$Cardio ~ train$Age + train$Height + train$Weight +
    train$AP_LO + train$Cholesterol + train$Smoke + train$Alcohol +
    train$Active, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8540  -1.1623   0.7285   1.0316   1.7063

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.9060219  0.3931186 -12.480  < 2e-16 ***
train$Age                 0.0900379  0.0041785  21.548  < 2e-16 ***
train$Height             -0.0098313  0.0019305  -5.093 3.53e-07 ***
train$Weight              0.0209445  0.0011494  18.223  < 2e-16 ***
train$AP_LO               0.0014703  0.0001766   8.325  < 2e-16 ***
train$CholesterolGroup 2  0.6675971  0.0339312  19.675  < 2e-16 ***
train$Smoke1             -0.1611096  0.0574932  -2.802  0.00507 **
train$Alcohol1           -0.1135297  0.0703914  -1.613  0.10678
train$Active1            -0.2092284  0.0368381  -5.680 1.35e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28377  on 20884  degrees of freedom
Residual deviance: 26751  on 20876  degrees of freedom
AIC: 26769
```

- Hypothesis test using Wald's test to drop *Alcohol*:

$$H_0 : \beta_7 = 0; H_a : \beta_7 \text{ is not zero}$$

After predictors Gender and Glucose were dropped from the model, in Figure 11, it can be seen that the p-value for *Alcohol* is **0.10678.** Therefore, with a confidence level of *0.05*, we fail to reject the null hypothesis and can remove the predictor from the model because it is not significant.

- Hypothesis test using Chi-Square Model Deviance to determine if the model is useful in predicting people who develop cardiovascular disease:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

$$H_a : \text{at least one of the coefficients in the null hypothesis is not zero}$$

Our logistic regression model is compared with the intercept only model with the process of finding the difference between the null deviance and residual deviance found in Figure 12. Using the chi-square model deviance to determine the p-value and the following function in R *1-pchisq(1623,7)*, the calculated p-value is **0**. Therefore, with a confidence

level of *0.05*, our data supports the claim that our logistic regression model is useful in estimating the log odds of developing cardiovascular disease.

Figure 12: Summary of Model without Gender, Glucose, and Alcohol

```
Call:
glm(formula = train$Cardio ~ train$Age + train$Height + train$Weight +
    train$AP_LO + train$Cholesterol + train$Smoke + train$Active,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8524  -1.1613   0.7287   1.0314   1.7067

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.8936545  0.3930366 -12.451  < 2e-16 ***
train$Age                 0.0901056  0.0041782  21.566  < 2e-16 ***
train$Height             -0.0099160  0.0019298  -5.138 2.77e-07 ***
train$Weight              0.0208921  0.0011487  18.188  < 2e-16 ***
train$AP_LO               0.0014699  0.0001765   8.327  < 2e-16 ***
train$CholesterolGroup 2  0.6646955  0.0338759  19.621  < 2e-16 ***
train$Smoke1             -0.1887472  0.0548588  -3.441  0.00058 ***
train$Active1            -0.2107440  0.0368235  -5.723 1.05e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28377  on 20884  degrees of freedom
Residual deviance: 26754  on 20877  degrees of freedom
AIC: 26770
```

After performing additional research to better understand the common factors of people that developed cardiovascular disease, we found BMI to be a prevalent component. We calculated and created a BMI predictor using the Weight and Height predictors with the hopes of strengthening the predictive ability of the model while also reducing the complexity. It was determined that the BMI predictor was not statistically significant while in the presence of the other predictors and therefore was not included in the model. We would like to perform similar procedures in the future to determine if there are better predictors for our model.

Results

The final model will help to explain if an individual has heart disease or not based on various predictors. The final model we found is:

log (odds) = -4.8936545 + 0.0901056*(Age) - 0.0099160*(Height) + 0.0094765*(Weight) + 0.0014699*(Diastolic blood pressure) + 0.6646955*(If that individual has high cholesterol) - 0.1887472*(If that individual smokes)- 0.2107440*(If that individual is active).

The three assumptions of a logistic regression model have been met with this final model:

- Observations are independent: All observations are not related or influenced by the measurements of other observations.
- Response is binary: The response is either 1 or 0 i.e. if an individual has cardiovascular disease or not.
- Log odds is a linear combination of a coefficient and predictors: The log odds of whether or not an individual has heart disease is based on a coefficient and seven predictors as shown by the model above.

Furthermore, the coefficients of the logistic regression model can be interpreted as follows for age and height. The estimated difference in log odds of someone developing cardiovascular disease is 0.0901056, per 1 year increase in age while all other variables are held constant. To further explain, the estimated difference in log odds of someone developing cardiovascular disease is -0.0099160, per 1 centimeter increase in height while all other variables are held constant. This method of interpretation can be used for all predictors.

To further understand real-life examples, we have identified three scenarios to predict the odds and probability of someone developing cardiovascular disease. The scenarios below include an "unhealthy" individual, a "healthy" individual, and a moderate example in between.

Scenario 1: An "unhealthy" individual

- Age is 62, height is 180cm, weight is 100 kg, diastolic bp of 97 mmHg, this individual does have high cholesterol, does smoke, and is not active

- When using this person's demographic and health information in our logistic regression model we receive the following:

  log(odds) = -4.8936545 + 0.0901056*62 - 0.0099160*180 + 0.0094765*100 + 0.0014699*97 + 0.6646955*1 - 0.1887472*1 - 0.2107440*0

- The estimated odds of this individual having cardiovascular disease is 1.606714, and the probability is 0.6163753*100 = 61.63753%.

```
# Scenario 1
# Unhealthy - 62 years old, 180 (5'9), 100, 97, 1, 1, 0
x = -4.8936545 + 0.0901056*62 - 0.0099160*180 + 0.0094765*100 + 0.0014699*97 + 0.6646955*1 - 0.1887472*1 - 0.2107440*0
odds = exp(x)
odds# 1.606714
prob = exp(x)/(1+exp(x))
prob #0.6163753
```

Scenario 2: A "healthy" individual

- Age is 35, height is 160cm, weight is 61kg, diastolic bp of 95 mmHg, does not have high cholesterol, does not smoke, and is active
- When using this person's demographic and health information in our logistic regression model we receive the following:

  log(odds) = -4.8936545 + 0.0901056*35 - 0.0099160*160 + 0.0094765*61 + 0.0014699*95 + 0.6646955*0 - 0.1887472*0 - 0.2107440*1

- The estimated odds of this individual having cardiovascular disease is 0.05963244, and the probability is 0.05627654*100 = 5.6276%, which is much lower than the unhealthy individual, as expected.

```
> # Healthy - 35 years old, 160, 135, 95
> x = -4.8936545 + 0.0901056*35 - 0.0099160*160 + 0.0094765*61 + 0.0014699*95 + 0.6646955*0 - 0.1887472*0 - 0.2107440*1
> odds = exp(x)
> odds#  0.05963244
[1] 0.05963244
> prob = exp(x)/(1+exp(x))
> prob #0.05627654
[1] 0.05627654
```

Scenario 3: A moderate individual

- Age is 70, height is 175cm, weight is 70kg, diastolic bp is 100mmHg, does not have high cholesterol, does not smoke, and is active
- When using this person's demographic and health information in our logistic regression model we receive the following:

  log(odds) = -4.8936545 + 0.0901056*70 - 0.0099160*175 + 0.0094765*70+ 0.0014699*100 + 0.6646955*0 - 0.1887472*0 - 0.2107440*1
- The estimated odds of this individual having cardiovascular disease is 1.320537, and the probability is 0.5690653*100 = 56.276%, which is slightly lower than the unhealthy individual in Scenario 1.

```
> # Scenario 3
> x = -4.8936545 + 0.0901056*70 - 0.0099160*175 + 0.0094765*70 + 0.0014699*100 + 0.6646955*0 - 0.1887472*0 - 0.2107440*1
> odds = exp(x)
> odds#  1.320537
[1] 1.320537
> prob = exp(x)/(1+exp(x))
> prob #0.5690653
[1] 0.5690653
```

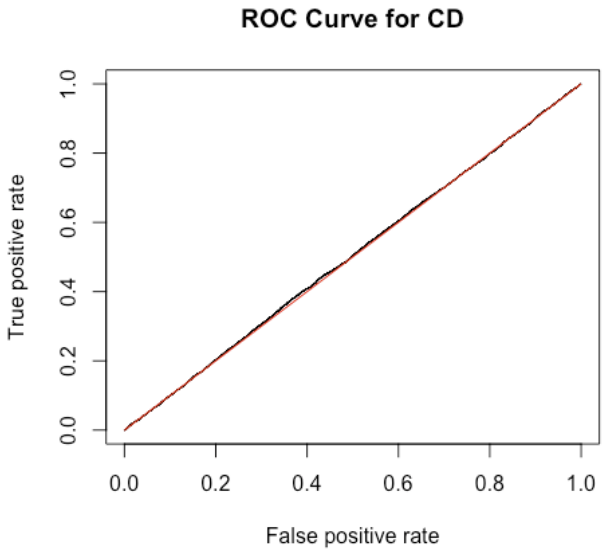
Additionally, we generated confidence intervals for age, diastolic blood pressure, and weight detailed below:

- **Age**: We have 95% confidence the odds of someone developing cardiovascular disease is multiplied by a value between (exp(0.08191633), exp(0.09829487)) = (1.085365, 1.103288) times for a unit increase of age, while all other variables are held constant.
- **Diastolic blood pressure**: We have 95% confidence the odds of someone developing cardiovascular disease is multiplied by a value between (exp(0.00112396), exp(0.00181584)) = (1.001125, 1.001817) times for a unit increase in diastolic blood pressure, while all other variables are held constant.

- **Weight**: We have 95% confidence the odds of someone developing cardiovascular disease is multiplied by a value between (exp(0.00845484), exp(0.01049716)) = (1.008491, 1.010552) times for a unit increase in weight, while all other variables are held constant.

In order to interpret how well our model does in classifying our data, we focused on the ROC curve and the confusion matrix. As seen by the ROC curve in Figure 13, our model performs just barely above random guessing with an AUC of 0.5085568. However, because this is a rare event, the overall error rate, the ROC curve, and the AUC could all be misleading, so we will need to focus on the confusion matrix as mentioned during the April 21, 2021 Live Session Section 2-26:42.

Figure 13: ROC Curve



We also decided to focus on the false negative rate because that is most relevant to this situation as it is worse to be told that you are negative when you actually do have heart disease, so we wanted to minimize this value.

We found the false negative and false positive rate with a threshold of 0.5 as seen in Figure 14:

FNR= False Negative/(False Negative + True Positive) = 3458/(3458+7827) = **0.3064245**
FPR= False Positive/(False Positive + True Negative) = 5626/(5626+2629) = **0.68152635**

Figure 14: Confusion Matrix with 0.5 threshold values

```
> table(test$Cardio, preds>0.5)

    FALSE TRUE
  0  2629 5626
  1  3458 7827
```

However, because we want to minimize the False Negative Rate and increase the False Positive Rate, we changed the threshold value to 0.4 as seen in Figure 15:

FNR= False Negative/(False Negative + True Positive) = 863/(863+10422) = **0.07647319**
FPR= False Positive/(False Positive + True Negative) = 7549/(7549+706) = **0.91447608**

Figure 15: Confusion Matrix with 0.4 threshold values

```
> table(test$Cardio, preds>0.4)

    FALSE  TRUE
  0   706  7549
  1   863 10422
```

Conclusion

Based on results of our data analysis, we conclude that there are seven predictors that are useful in predicting if someone will develop cardiovascular disease or not. We would like to take additional next steps to further enhance our logistic regression model by beginning with reevaluating the data source. We believe a better dataset can be identified to explore additional risks and parameters that are more appropriately related to cardiovascular disease. Within the current dataset, the person with the highest age is 64 years old and we have seen that cardiovascular disease is more prevalent in people older, more specifically older than 65. Therefore, a dataset that includes people in this age group may strengthen our model. Additionally, we are interested in performing more research on incorporating interaction terms into the model.

Sources

Cardiovascular Disease dataset. (2019, January 20). Kaggle.

https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Heart Health and Aging. (2021). National Institute on Aging.

https://www.nia.nih.gov/health/heart-health-and-aging

Know Your Risk for Heart Disease | cdc.gov. (2019, December 9). Centers for Disease Control and Prevention.

https://www.cdc.gov/heartdisease/risk_factors.htm#:%7E:text=About%20half%20of%20all%20Americans,the%20factors%20you%20can%20control.

Understanding Blood Pressure Readings. (2021). Www.Heart.Org.

https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings