Applying Classification Methods to Predict Heart Disease and Using Linear Regression to Identify Associations Between Physiological Indicators

Final Project

STA 141A Fall 2021

Nancy Hernandez (nalhernandez@ucdavis.edu), Sam Van Gorden (spvangorden@ucdavis.edu), Robin Boparai (rboparai@ucdavis.edu)

Contributor	Contribution
Nancy	Answering research questions, Inferences, Model setup, Editing, Background, Prediction function, Commenting, Data cleaning
Sam	Answering research questions, Inferences, Model setup, Code design, Helper functions, Data cleaning
Robin	Answering research questions, Inferences, Model setup, RMD editing, Introduction, Conclusion, Assist with data cleaning

Note: We created four functions for our project. Three functions are helper functions (included in Appendix B) that aided our analysis, and the fourth function is a prediction function (included in Appendix A) that can provide a predicted diagnosis of heart disease given one's performance on relevant physiological indicators.

Introduction

Heart disease affects a large proportion of the U.S. population and imposes significant health and economic burdens on those affected. About 27.6 million adults in the United States have been diagnosed with heart disease (Benjamin et al., 2018). According to estimates from 2003, about 50 million adults in the United States have hypertension, and about 62 million people are thought to have cardiovascular disease (Nabel, 2003). Cardiovascular disease risk is also believed to differ based on some demographic characteristics (Cooper, 2001). In order to understand risk factors associated with heart disease, we performed an analysis of one of the largest publicly available data sets for heart disease and related risk factors.

Heart Disease Data

The purpose of this project is to utilize one of the largest heart disease data sets publicly available to identify possible risk factors that predict heart disease. Heart disease includes ailments that involve cardiovascular pathology (Benjamin et al., 2018).

The data set used in this analysis was obtained from the UCI Machine Learning Repository and includes data collected by the Hungarian Institute of Cardiology, University Hospitals in Zurich, Switzerland and Basel, Switzerland, the V.A. Medical Center in Long Beach, and the Cleveland Clinic Foundation. In total, the data set includes observations on key physiologic indicators for 918 patients.

The data set used in our analysis includes five continuous variables: age, resting blood pressure, serum cholesterol level, maximum heart rate, and depression in the ST wave induced by exercise relative to rest. The data set includes observations for male and female adults between the ages of 31 and 77 years. Resting blood pressure (mm Hg) was collected during periods of inactivity. Serum cholesterol level (mm/dl) refers to the concentration of cholesterol in one's blood and is determined via a blood test. Maximum heart rate (beats per minute) was collected during periods of exertion. Depression in the ST segment (also called "oldpeak") refers to the change in the interval between ventricular depolarization and ventricular repolarization caused by exercise.

The data set also includes seven categorical variables: sex, chest pain type, fasting blood sugar status, resting electrocardiogram (ECG) results, exercise angina status, slope of the peak exercise ST segment, and heart disease diagnosis. Chest pain type includes the levels "typical angina (TA)", "atypical angina (ATA)", "non-anginal pain (NAP)", and "asymptomatic (ASY)". In the data set, an individual is classified as having a high fasting blood sugar if his or her blood sugar is greater than 120 mg/dl. Resting ECG results includes the levels "Normal", "ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)", and "LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria". Slope of the peak exercise ST segment includes the levels "up", "flat", and "down". Exercise angina status and heart disease diagnosis are both binary variables.

The dataset included some missing values for serum cholesterol; serum cholesterol data was not collected for all research subjects. Thus, we removed observations for whom serum cholesterol data was missing from our dataset. After removing observations for which data was missing, our dataset had 746 observations in total. Moreover, we subset the data based on key demographic characteristics and physiological indicators to focus our analyses on specific subgroups when addressing some of our research questions.

Research Question I

Using linear regression and model selection to identify significant predictors of exercise-induced depression in the ECG ST segment among men with exercise angina

Exercise angina occurs when an individual is preforming some form of physical activity, and the heart muscle is not receiving enough oxygen. One of the most common causes for males is the narrowing of arteries caused by fatty deposits. This can occur when one has high cholesterol, and cardiovascular pathology is associated with heart rate abnormalities, resting blood pressure, and age. We can create a model that determines whether or not these risk factors are associated with exercise-induced depression in the ECG ST segment. The exercised-induced depression in the ECG ST segment is important to assess in a clinical context because it is an important physiological indicator that can improve the clinical information derived from an exercise stress test.

Thus, we pose the following research question: *Do resting blood pressure, maximum heart rate, and age have significant effects on exercise-induced depression of the ECG ST segment among men with exercise angina?*

Model Set-Up: Full model: `Oldpeak ~ MaxHR + RestingBP + Age`

From the normal probability (Normal Q-Q) plot in Figure 1, the assumption of normality holds because most dots are close to or on the theoretical line. However, some outliers exist; these outliers include observation 600, 167, and 620.

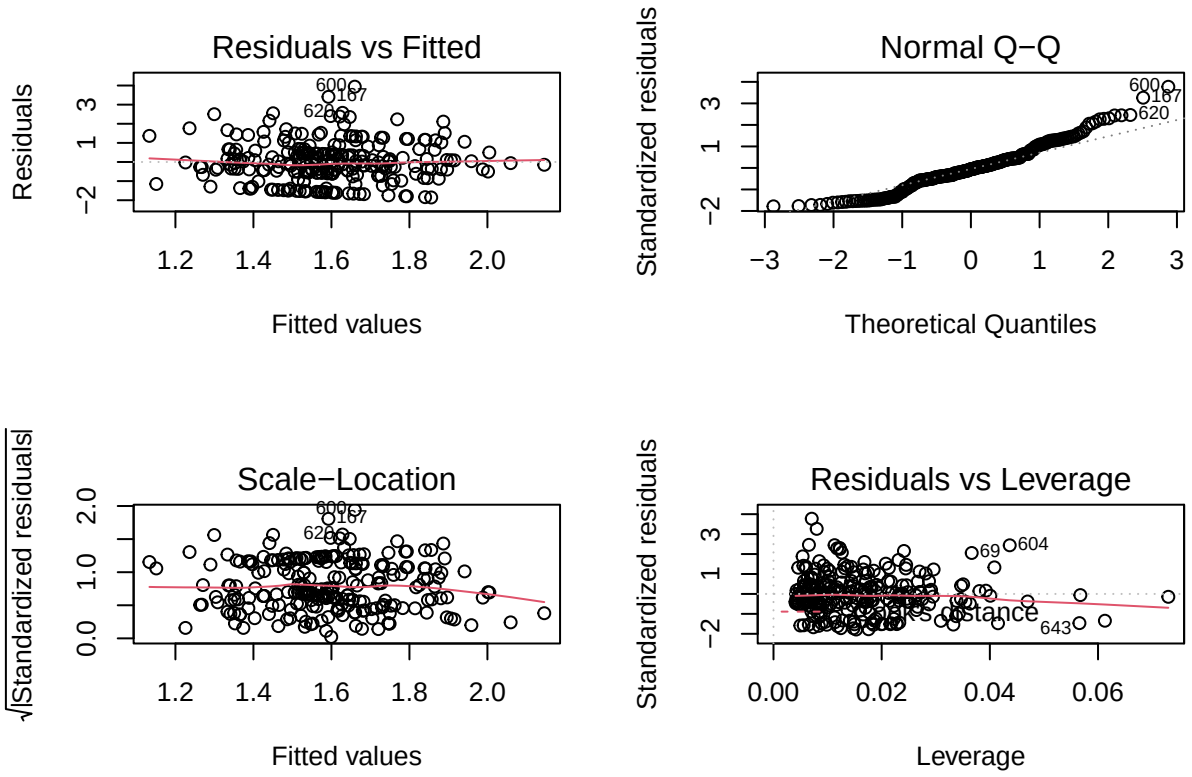

Figure 1: Residual/diagnostic plots for the full linear regression model.

From observing the Residuals vs Fitted plot in Figure 1, there appears to be no discernible pattern in the relationship between the residuals and the fitted values. In addition, the residuals roughly form a horizontal band around the 0 line. Therefore, the assumption of equal variance holds. Using the Residuals vs Leverage plot, we know that there may be some high leverage points in the data set because there are points where their predictor x-value is far greater than the predictor x-values of the majority of the points.

Summary of the full model:

The full model has an $R^2 = 0.014$, suggesting that the model explains only a small proportion of the variability in the response variable. The residual standard error of the model is 1.048. When performing an F-test to evaluate the model, a p-value of 0.091 is obtained, suggesting that the model does not provide a better fit for the data than a model that includes no predictors. Thus, we use model selection techniques to improve our model.

Model Selection:

The stepwise selection shows that the final model should not include the variables maximum heart rate and age. It can also be found that the predictors maximum heart rate and age have the smallest AIC statistics, which suggests that maximum heart rate and age can be dropped from the model. Therefore, in the following analysis, we use the reduced model.

Reduced model: `Oldpeak ~ RestingBP`

The residual plot in Figure 2 show that there is no obvious pattern in the residuals and that the residuals appear to be evenly distributed around the zero line. In addition, the normal probability plot in Figure 2 is closely associated with the dotted line; it departs only slightly on both head and tail. Thus, the error terms of the this regression model appear to be approximately normal and have constant variance.

Summary of the reduced model: Using a linear regression model where exercise-induced depression in the ST segment of an ECG reading is used as the response variable and resting blood pressure is used as the predictor variable, resting blood pressure is found to be a significant predictor among males with exercise angina. We can look at the p-value of the regression coefficient in order to determine whether our variable is significant. The estimated regression coefficient for the variable resting blood pressure is statistically significant with a p-value of 0.015. However, the adjusted R^2 value of the model is only 0.0201, suggesting that the predictor resting blood pressure does not explain most of the variability in the the exercise-induced depression of the ST segment among males with exercise angina. Further research questions are addressed below to identify

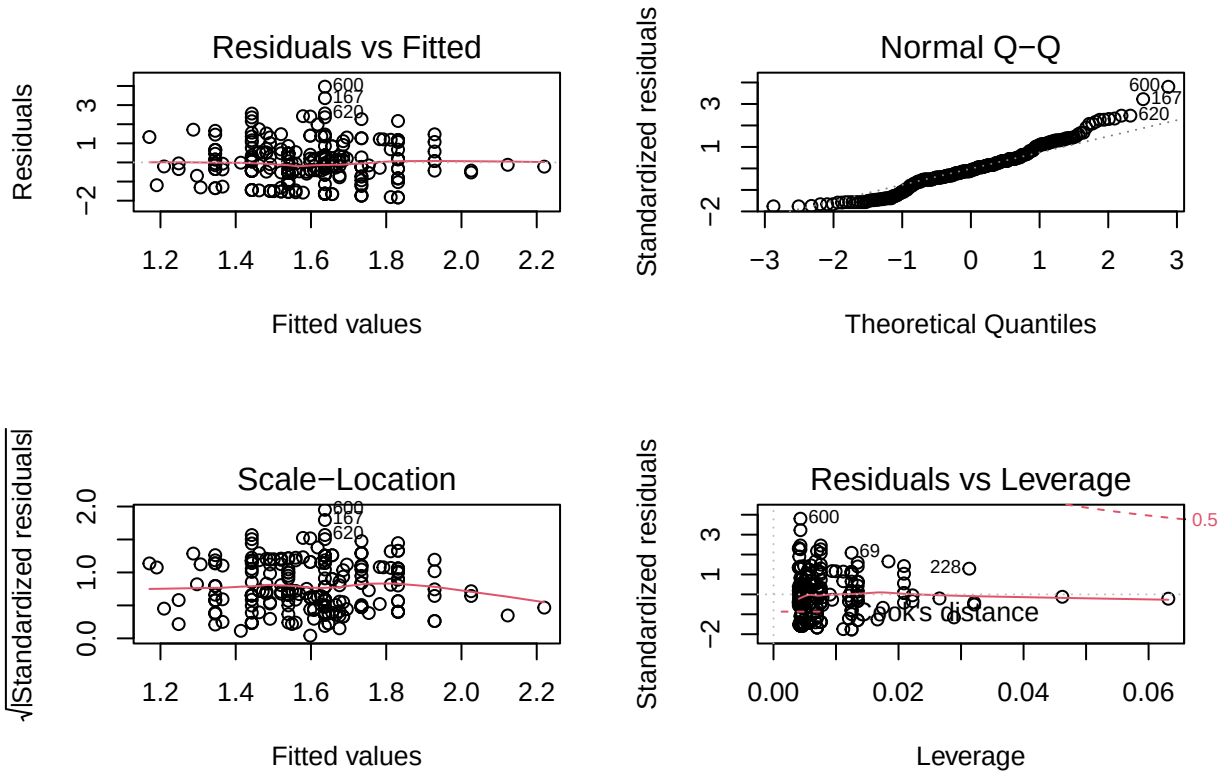

Figure 2: Residual/diagnostic plots for the reduced linear regression model.

risk factors that have an effect on one's probability of developing heart disease, and investigations of associations between physiological indicators included in this dataset are discussed below.

Research Question II

Identifying risk factors associated with heart disease across different age groups using logistic regression

As an individual ages, the heart and blood vessels age as well. The vessels begin to deteriorate slowly with age in both males and females. There are age-related risk factors that are present as the body starts to age, so they might not be prevalent until a later stage in life. Older adults especially of age 65 or older have been seen to have a higher risk of developing heart disease and other complications when compared to younger adults.

Thus, we present the following research question: *Is the relationship between heart disease and associated risk factors different across age groups?*

Age Groups:

Group 1: 31-49 Group 2: 50-56 Group 3: 57-61 Group 4: 62+

Predictor Coefficients:

Group 1 :

	Estimate	Std. Error	z value	Pr(>|z|)
SexM	2.471609	0.5936545	4.163379	0.0000314
ChestPainTypeATA	-2.716756	0.6801838	-3.994150	0.0000649
ChestPainTypeNAP	-1.537624	0.5412604	-2.840820	0.0044998
ST_SlopeUp	-2.992931	1.1341819	-2.638846	0.0083189

Group 2 :

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.271838	2.6196106	-2.775923	0.0055045
SexM	2.164845	0.6639852	3.260381	0.0011126
ChestPainTypeNAP	-1.746836	0.6371090	-2.741816	0.0061101
ChestPainTypeTA	-3.383885	1.3389921	-2.527188	0.0114980
ExerciseAnginaY	1.082579	0.5451452	1.985854	0.0470495
ST_SlopeUp	-2.515508	0.9622400	-2.614221	0.0089431

Group 3 :

	Estimate	Std. Error	z value	Pr(>|z|)
RestingECGNormal	-1.124957	0.5472912	-2.055500	0.0398308
ExerciseAnginaY	1.445406	0.5212510	2.772956	0.0055550

Group 4 :

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.0194000	1.8138140	-2.215994	0.0266919
SexM	2.8093268	0.7947713	3.534761	0.0004081
ChestPainTypeTA	-2.8404332	1.0268611	-2.766132	0.0056726
ExerciseAnginaY	1.3689122	0.6763696	2.023911	0.0429793
Oldpeak	0.8512758	0.3906874	2.178918	0.0293378

First we split the data into four age groups based on the four quartiles of age data. Then, we fit each group to a logistic regression model with presence of heart disease as the response variable and the remaining variables as predictors. After this, we pass each model into the stepAIC function to determine which predictors should be kept for each group. Finally, we display all of the remaining predictor coefficients for predictors significant at the 0.05 level of significance.

The following are the significant predictors of heart disease for each age group, ordered from greatest to least amount of influence (i.e. absolute value of coefficient):

Age 31-49: ST_Slope (Up), ChestPainType (ATA), Sex (M), ChestPainType (NAP)

Age 50-56: ChestPainType (NAP), ChestPainType (TA), ST_Slope (Up), Sex (M), ExerciseAngina (Y)

Age 57-61: RestingECG (Normal), ExerciseAngina (Y)

Age 62+: ChestPainType (TA), Sex (M), Oldpeak

Based on these observations it appears that three types of chest pain (atypical angina, non-anginal pain, and typical angina) are significant and influential in determining heart disease. Strangely, these predictors all have negative coefficients which indicates that the presence of these types of chest pain predicts no heart disease. This seems unusual because it is common to think of chest pain as being a precursor to various heart conditions. Because of this, more research may be necessary or perhaps a different sort of analysis should be done to verify this counterintuitive result. The absence of chest pain (i.e. ChestPainType = ASY) does not appear in the logistic regression, so apparently not having chest pain is not an influential factor in predicting the presence or absence of heart disease. Sex (Male) appears with a positively-valued coefficient in three of the four age groups as well. This seems to make sense as it is common to think of men as being more prone to heart disease than women. The presence of exercise angina appears in the three older-age groups, again with positive coefficient, though lower than the Sex (Male) coefficient. This indicates that the presence of exercise angina is an influential predictor of heart disease in older people, though probably not as influential as being male.

Upward sloping of the peak exercise ST segment shows up with a relatively high negative coefficient in the two younger group models, so it may be an influential predictor of not having heart disease but only for younger individuals. Normal resting ECG shows up with a negative coefficient for Group 3 only and oldpeak with a positive coefficient for Group 4 only.

In general, it seems that the predictors that show up in multiple age groups (chest pain, sex - male) tend to have higher influence than those predictors that only show up in one age group (oldpeak, normal resting ECG). Based on this observation, we can infer that the strongest predictors of the presence or absence of heart disease do not depend on age. Another interesting result of this analysis is that all selected predictors were categorical variables except for oldpeak, which also had the lowest predictive power of all predictors.

The plot in Figure 3 confirms that chest pain is more common among research subjects without heart disease in the dataset. The plot shows the frequency of patients with each chest pain type depending on whether or not they have been diagnosed with a heart disease. The plot in Figure 4 illustrates the age range of the subjects included in the dataset.

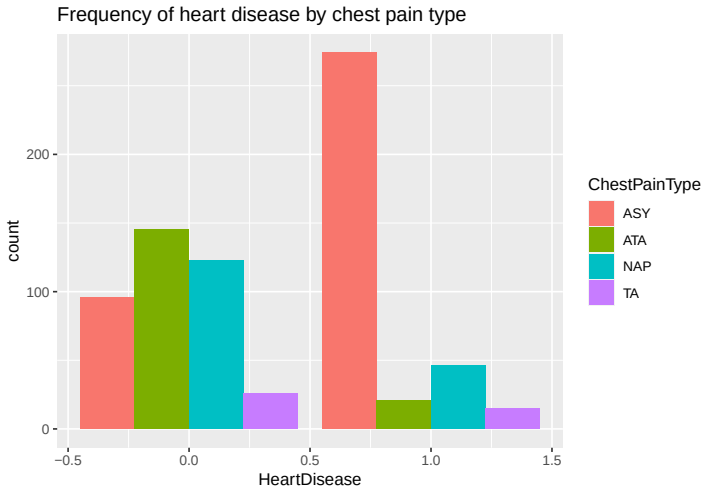

Figure 3: The histogram shows the number of patients with and without heart disease across levels of chest pain type (chest pain atypical angina (ATA), typical angina (TA), asymptomatic (ASY), and non-anginal pain (NAP)). There appears to be a greater proportion of research subjects with no chest pain among those with heart disease.

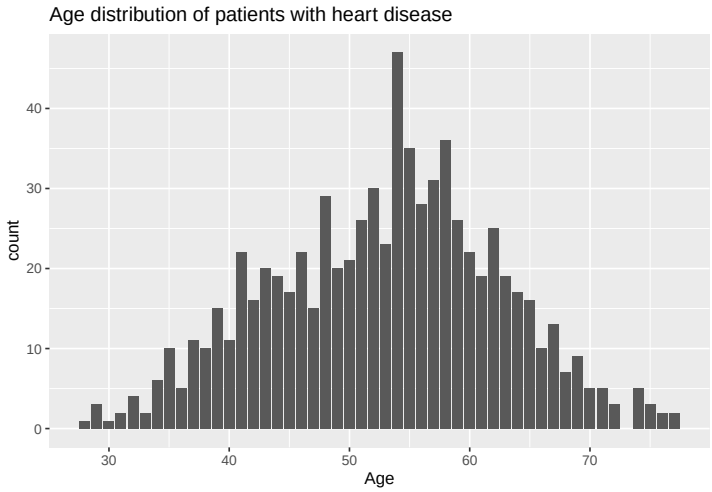

Figure 4: The histogram shows the age distribution of the research subjects. Age appears to be approximately normal.

Research Question III

Comparing classification methods (logistic regression, LDA, and k-NN)

We are interested in which type of modeling works best if we are to use our dataset to predict future occurrences of heart disease. For this reason, we would like to compare the results of logistic regression (logit), linear discriminant analysis (LDA), and k-Nearest Neighbors (k-NN) on our dataset. We would also like to compare the results of these models on the full subset of parameters as well as a partial subset of potentially more influential parameters to avoid overfitting. For the partial model, we will use the predictors from Research Question II that were found in three or more of the four age groups. It may also be useful to perform F-tests on the two logistic regression models as well as a model generated using stepAIC which optimizes the number and choice of parameters by minimizing the AIC formula.

Therefore, we can ask the following question: *Which classification algorithm provides the lowest misclassification error rate for the full and partial datasets?*

Full model results:

Logistic Regression: Misclassification Error Rate = 0.1344

	Predicted Heart Disease	Predicted No Heart Disease
Observed Heart Disease	78	16
Observed No Heart Disease	9	83

Linear Discriminant Analysis: Misclassification Error Rate = 0.1452

	Predicted Heart Disease	Predicted No Heart Disease
Observed Heart Disease	72	20
Observed No Heart Disease	7	87

k-Nearest Neighbors: Misclassification Error Rate = 0.3495

	Predicted Heart Disease	Predicted No Heart Disease
Observed Heart Disease	76	28
Observed No Heart Disease	37	45

Partial model from RQ2 results:

Logistic Regression: Misclassification Error Rate = 0.2151

	Predicted Heart Disease	Predicted No Heart Disease
Observed Heart Disease	88	18
Observed No Heart Disease	22	58

Linear Discriminant Analysis: Misclassification Error Rate = 0.2366

	Predicted Heart Disease	Predicted No Heart Disease
Observed Heart Disease	74	26
Observed No Heart Disease	18	68

k-Nearest Neighbors: Misclassification Error Rate = 0.2366

	Predicted Heart Disease	Predicted No Heart Disease
Observed Heart Disease	73	27
Observed No Heart Disease	17	69

We ran the classifications several times with different seeds to randomize the test and training data used by each classification method. The results obtained were similar for each iteration - LDA and logistic regression provided lower misclassification error rates for the full model and k-NN underperformed both for the full model and was similar to the misclassification error rates for LDA and logistic regression for the partial model. From this we can conclude that the quantitative predictors are likely normally distributed, since only the full model contains quantitative predictors and LDA and logistic regression perform better for normally distributed predictors.

Among all of the iterations we performed, the misclassification rates for LDA and logistic regression were lower (or similar for partial model) than for k-NN. For LDA and logistic regression, the misclassification error rates were lower for the full model than they were for the partial model. Based on these results it seems it is preferable to use either LDA or logistic regression on the full model, or another model with a different subset of predictors than those selected from RQ2, for further classification efforts.

We experimented with different values of k for k-NN and found that k=12 produced the best results. It is recommended to use a value close to the square root of the number of test data points (186 in our case) which would be between 13 and 14. Therefore we started from k=14 and worked down to k=12 which seemed to provide better results (lower misclassification error rate).

Research Question IV

Generating a model to predict heart disease among individuals with atypical angina using physiological indicators

Atypical angina occurs when one has chest pain that does not have a common diagnosis and may not be typical indicators of chest pain. This could mean that one may have unusual symptoms or various symptoms that may not fit into a single diagnosis. Since this type of chest pain is difficult to understand, we can try to create a model that may determine the probability of someone having heart disease with atypical angina.

We will address the following research question: *Given one's resting blood pressure, serum cholesterol level, and maximum heart rate, what is the probability that an individual with atypical angina has heart disease?*

The research question can be interpreted as a classification problem in which the categorical outcome variable is heart disease and the qualitative predictor variables are resting blood pressure, serum cholesterol level, and maximum heart rate. In order to evaluate the probability of developing heart disease among those with atypical angina specifically, we subset the dataset to include only those with atypical angina. Based on our analysis from Research Question III, it is appropriate to utilize a logistic regression model in this context. We will focus our analysis on the model below.

Model Set-Up : Full model: `HeartDisease ~ RestingBP + Cholestrol + MaxHR`

The following logistic regression model was used to predict the probability of developing heart disease given one's resting blood pressure, cholesterol, and maximum heart rate.

$$Pr(Y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}$$

X_1 is one's resting blood pressure, X_2 is one's serum cholesterol level, and X_3 is one's maximum heart rate.

We can split the data, so we can have 50 randomly sampled observations as test data and the rest of the data be training data. Now we create the multiple logistic regression model.

	Predicted Heart Disease	Predicted No Heart Disease
Observed Heart Disease	44	1
Observed No Heart Disease	5	0

A confusion table is created to compare the true values of the test data versus the predicted outcome of predicting heart disease using the training data.

Test Accuracy

We can visualize some characteristics by plotting each variable against each other. (See Figure 5). In each of the three plots, we can see how it is difficult to discriminate between having heart disease and not having heart disease. This is because they are similar in terms of max heart rate, resting blood pressure, and cholesterol. These plots illustrate whether or not it would be difficult to tell apart patients who have heart disease from those who do not. From our data, all the characteristics are very close so its going to be very difficult, which is confirmed when solving for the type II error and true positive rate. If the variables in the training data are compared to one another, we see that there is not much difference between observations with those who have heart disease and those who do not. This will affect the accuracy of our result because the model will be more likely to mix up the observations in the test data.

Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = 0.88$

The model accuracy for testing the probability that an individual with atypical angina has heart disease given one's resting blood pressure, serum cholesterol level, and maximum heart rate, is 88% which is high. This test accuracy rate can be misleading especially after looking at the plots comparing the variables, so we must consider looking at type I and II errors as well as sensitivity and specificity.

Type I error : $\alpha = \frac{FP}{N} = \frac{FP}{TN+FP} = 0.0222$

The type I error calculates the false positive rate which in our case is about 2.22%. This means that for every 100 people who do not have heart disease, about 2 people will be falsely told they have heart disease.

Type II error : $\beta = \frac{FN}{P} = \frac{FN}{FN+TP} = 1$

The type II error calculates the false negative rate which in our case is 100%. This means that for every 100 people who do have heart disease, 100 people will be falsely told they do not have heart disease. This is not very accurate and can be serious if the individual is not treated appropriately.

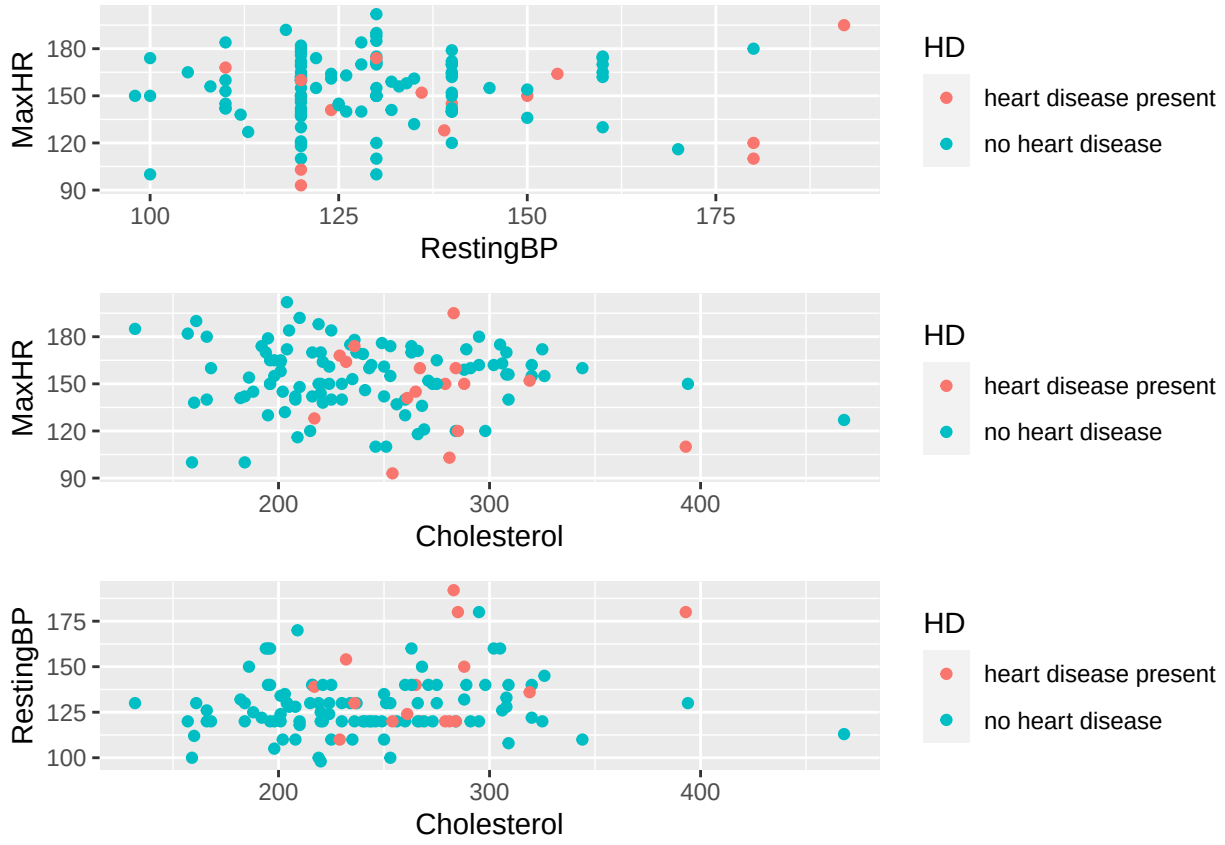

Figure 5: Scatter plots for selected predictors included in the logistic regression model.

Sensitivity (True Positive Rate) : $TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = 0$

Sensitivity calculates the true positive rate which in our case is 0%. This means that for every 100 people who do have heart disease, none of the people will be positively told they have heart disease. This rate of 0% is extremely poor, and tells us that the test is not able to determine which patient has heart disease which is a major flaw.

Specificity (True Negative Rate) : $TNR = \frac{TN}{N} = \frac{TN}{TN+FP} = 0.9778$

Specificity calculates the true negative rate which in our case is about 97.78%. This means that for every 100 people who do not have heart disease, about 98 of them will be positively told they do not have heart disease. The rate is fairly high, so there is a likely chance individuals with no heart diesease will be diagnosed accurately.

Based on these rates, the model works best calculating those who do not have heart disease versus those who do have heart disease. The true negative rate was about 97.78%. This is consistent with the accuracy rate. A major reason as to why we may have received such a low true positive rate of 0% is that our test data did not include enough observations with heart disease and those that were included could have been similar to observations that did not have heart disease. Also, the subjects with heart disease and the subjects without heart disease perform similarly on some of the predictor variables our dataset; this is inherent in our dataset. Thus, in predicting whether or not an individual had heart disease, there was not enough information to show that they do in fact have heart disease. This could mean that atypical angina may not be a significant contributor to heart disease as previously believed, or that our test data should have included more observations that did have heart disease.

Conclusion

Our findings indicate that some physiological indicators (exercise angina, age, oldpeak, chest pain, etc.) are associated with one's probability of being diagnosed with heart disease. However, our findings were counterintuitive in some areas. For example, we found that chest pain is more common among research subjects without heart disease in this dataset. This counterintuitive finding may be attributed to the fact that chest pain can be misdiagnosed due to subjective diagnostic criteria or a misidentification of symptoms. Moreover, even though chest pain is believed to predict heart disease, our analysis of the UCI dataset obtained via Kaggle suggests that chest pain may not indicate a greater probability of being diagnosed with heart disease. We are limited in the interpretations we can make from our dataset since the sample is not representative. The

sample only included individuals living in a few cities in CA, OH, and Europe); the subpopulation of patients in the dataset may not be representative of the entire population. Lastly, our findings demonstrated that one's resting blood pressure is a significant classifier for heart disease among individuals with atypical angina. Further analyses must be performed to better understand how the risk factors addressed in this analysis are associated with heart disease in studies that include samples that are representative of the general population.

Prediction Function

There are four types of chest pain atypical angina (ATA), typical angina (TA), asymptomatic (ASY), and non-anginal pain (NAP). We can create a function to predict whether one is at a high risk for developing heart disease given his or her type of chest pain, resting blood pressure, cholesterol, and max heart heart. (Depending on the type of chest pain that occurs, the accuracy rate of predicting whether one may have heart disease varies. This is because some chest pain types have a higher chance of showing future heart disease concerns versus others.) Please see Appendix A for the full prediction function.

Now that we have the function, we can input an example of an individual's data and determine whether they have a high risk of developing heart disease.

Predict whether a person has a high risk of developing heart disease and their probability if their chest pain type is atypical, their resting blood pressure is 160 mmHg, their cholesterol is 389mm/dl, and their maximum heart rate is 140 beats per minute.

```
predict_hd(ChestPain_type = "ATA",
           resting_bp = 160,
           chol = 389,
           max_hr = 140) # example
```

```
## $Probabilty
##        1
## 0.389045
##
## $Prediction
##                                                          1
## "May not have a high risk for developing heart disease"
```

References:

1. Benjamin, E. J., et al. (2018). Heart disease and stroke statistics—2018 Update: A report from the American Heart Association. *Circulation*, 137(12), 67-492.

2. Cooper, R. S. (2001). Social inequality, ethnicity and cardiovascular disease. *International journal of epidemiology*, 30, 48-52.

3. Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer, New York.

4. Heart Disease Data Set. *UCI Machine Learning Repository*. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D., University Hospital, Zurich, Switzerland: William Steinbrunn, M.D., University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D., 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

5. Nabel, E. G. (2003). Cardiovascular disease. *New England Journal of Medicine*, 349(1), 60-72.1.

Appendix A: Prediction Function

```
# ----- Prediction Function -----
predict_hd = function(ChestPain_type, resting_bp, chol, max_hr)
{
 # depending on the type of chest pain one has, the function will use the data
  # set of that specific chest pain type

  if (ChestPain_type == "ASY")
  {
    training_data = heart_data[heart_data$ChestPainType == "ASY",]
    # use only data with ChestPainType ASY
  }

  if (ChestPain_type == "ATA")
  {
    training_data = heart_data[heart_data$ChestPainType == "ATA",]
    # use only data with ChestPainType ATA
  }

  if (ChestPain_type == "NAP")
  {
    training_data = heart_data[heart_data$ChestPainType == "NAP",]
    # use only data with ChestPainType NAP
  }

  if (ChestPain_type == "TA")
  {
    training_data = heart_data[heart_data$ChestPainType == "TA",]
    # use only data with ChestPainType TA
  }

  test_data = tibble(RestingBP = resting_bp, Cholesterol = chol, MaxHR = max_hr)
  # set input data as test data

  glm_fit = glm(HeartDisease ~ RestingBP + Cholesterol + MaxHR,
                data = training_data)
  # create logistic regression model using the 3 predictors

  probability = predict(glm_fit, test_data, type = "response")
  # probability of the person developing heart disease

  prediction = ifelse(predict(glm_fit, test_data, type = "response") < 0.5,
                      "May not have a high risk for developing heart disease",
                      "May have a high risk for developing heart disease")
  # prediction of whether or not the person has heart disease

  return(list(Probabilty = probability, Prediction = prediction))
  #returns probability and prediction of having heart disease or not
}
```

Appendix B: Helper Functions

```r
## Helper Functions
## Creates a logistical regression model for the given data and family
## Returns list containing model, training data, and test data respectively
create_logit_model <- function(data, family = "gaussian") {
  data.copy <- data
  data.copy$id <- 1:nrow(data.copy)
  data.train <- dplyr::sample_frac(data.copy, .75)
  data.test  <- dplyr::anti_join(data.copy, data.train, by = 'id')
  data.model <- glm(HeartDisease ~ ., family = family, data = data.train)

  return(list(data.model, data.train, data.test))
}

## Creates a linear discriminant analysis model for the given data
## Returns list containing model, training data, and test data respectively
create_lda_model <- function(data) {
  data.copy <- data
  data.copy$id <- 1:nrow(data.copy)
  data.train <- dplyr::sample_frac(data.copy, .75)
  data.test  <- dplyr::anti_join(data.copy, data.train, by = 'id')
  data.model <- lda(HeartDisease ~ ., data = data.train)

  return(list(data.model, data.train, data.test))
}

## Creates a k-nearest neighbors model for the given data, predictors, observed value, and k value
## Returns list containing model, training data, and test data respectively
create_knn_model <- function(data, predictors, observed, k = 1) {
  data.copy <- data
  data.copy$id <- 1:nrow(data.copy)
  data.train <- dplyr::sample_frac(data.copy, .75)
  data.test  <- dplyr::anti_join(data.copy, data.train, by = 'id')

  # Convert non-numeric columns to numeric
  data.train <- dplyr::mutate_if(data.train, is.character, as.factor)
  data.test <- dplyr::mutate_if(data.test, is.character, as.factor)
  data.train <- dplyr::mutate_all(data.train, as.numeric)
  data.test <- dplyr::mutate_all(data.test, as.numeric)

  data.model <- knn(as.matrix(data.train[predictors]),
                    as.matrix(data.test[predictors]),
                    as.matrix(data.train[[observed]]), k = k)

  return(list(data.model, data.train, data.test))
}
```

Appendix C: Quantitative predictor variables

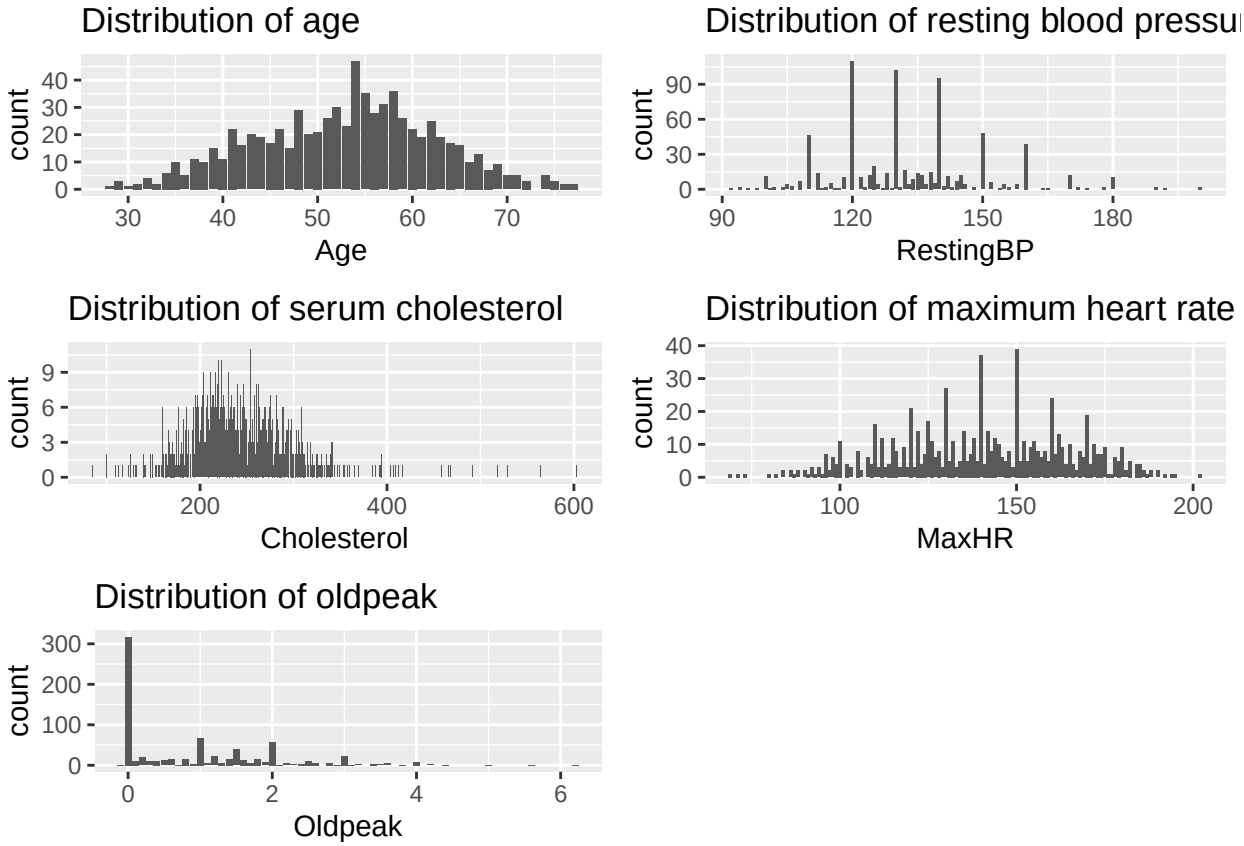

Figure 6: Histograms of quantitative predictor variables

Based on the visualizations in Figure 6, it appears that age and maximum heart rate are somewhat normally distributed. Resting blood pressure and serum cholesterol appear to have outliers. These outliers may affect the inferences from some of the analytical methods that we employed. Oldpeak appears to be skewed. "Oldpeak" refers to the change in the ECG interval between ventricular depolarization and ventricular repolarization caused by exercise; an individual can have a value of 0 for the variable oldpeak if they exhibit no change in the ECG interval between ventricular depolarization and ventricular repolarization caused by exercise.

Appendix D: Full Project Code

```
## Initialization
library(tidyverse)
library(MASS)
library(ggplot2)
library(gridExtra)
library(class)
heart_data <- read.csv("heart.csv")
# obtain our data set and save it to our global environment
heart_data = dplyr::filter(heart_data, Cholesterol != 0) # clean data
str(heart_data)
# display the structure of our data, outputting information about the variables
# ----- Research Question I -----
rq1_dataM = heart_data[heart_data$Sex == "M",] # Subset male data
rq1_dataY = rq1_dataM[rq1_dataM$ExerciseAngina == "Y",]
# Use only data where exercise angina is present

rq1_lmY = lm(Oldpeak ~ MaxHR + RestingBP + Age, data = rq1_dataY)
# Create a multiple linear Regression
summary(rq1_lmY)
rq1_dataN = rq1_dataM[rq1_dataM$ExerciseAngina == "N",]
# use only data where exercise angina is not present

rq1_lmN = lm(Oldpeak ~ RestingBP + MaxHR + Age, data = rq1_dataN)
# Create a multiple linear regression
summary(rq1_lmN)
par(mfrow = c(2,2))
plot(rq1_lmY)
summary(rq1_lmY)
# ------- model selection ------
stepAIC(rq1_lmY, direction ="both")
# In the last step, r selects the optimal model
rq1.selected_model <- lm(formula = Oldpeak ~ RestingBP, data = rq1_dataY)
par(mfrow = c(2,2))
plot(rq1.selected_model)
summary(rq1.selected_model)
# ----- Research Question II -----
# Break down of heart disease by age
rq2.data <- dplyr::filter(heart_data, HeartDisease == 1)
age.quantiles <- fivenum(rq2.data$Age)

# Print age groups
cat("Age Groups:")
cat("\n\n")
for (i in 1:3) {
  cat(paste("Group ", i, ": ", age.quantiles[i], "-", age.quantiles[i+1] - 1,
            sep = ""))
  cat("\n")
}
cat(paste("Group 4: ", age.quantiles[4], "+", sep = ""))
cat("\n\n")

rq2.data <- heart_data
rq2.data$AgeGroup <- factor(numeric(nrow(rq2.data)),
                            levels = 1:(length(age.quantiles)-1))

#create a new column called AgeGroup and set all data to its associated age group
rq2.data[rq2.data$Age <= age.quantiles[2], "AgeGroup"] <- 1
```

```
rq2.data[rq2.data$Age > age.quantiles[2] & rq2.data$Age <= age.quantiles[3],
         "AgeGroup"] <- 2
rq2.data[rq2.data$Age > age.quantiles[3] & rq2.data$Age <= age.quantiles[4],
         "AgeGroup"] <- 3
rq2.data[rq2.data$Age > age.quantiles[4], "AgeGroup"] <- 4

rq2.datasets <- dplyr::group_split(rq2.data, AgeGroup)
rq2.datasets <- lapply(rq2.datasets, subset, select = -c(Age, AgeGroup))
rq2.models <- list()

cat("Predictor Coefficients:")
cat("\n\n")

# Create model for each of the four age groups created
i <- 1
for (dataset in rq2.datasets) {
  rq2.models[[i]] <- glm(HeartDisease ~ ., data = dataset, family = binomial)
  rq2.models[[i]] <- stepAIC(rq2.models[[i]], direction = "both", trace = FALSE)

  # Only look at coefficients that are significant at the 0.05 level
  coeffs <- summary(rq2.models[[i]])$coefficients
  coeffs <- coeffs[coeffs,"Pr(>|z|)"] < 0.05,]

  cat(paste("Group ", i, ":"))
  print(kable(coeffs, col.names = colnames(coeffs)))
  cat("\n\n")

  i <- i + 1
}
ggplot(data = heart_data, aes(HeartDisease, ..count..)) +
  geom_bar(aes(fill = ChestPainType), position = "dodge") +
  ggtitle("Frequency of heart disease by chest pain type")
ggplot(data = rq2.data, aes(Age)) +
  geom_bar() +
  ggtitle("Age distribution of patients with heart disease")
# ----- Research Question III -----
set.seed(22)
for (i in 1:2) {
  if (i == 1) {
    rq3.data <- heart_data
    cat("Full model results:")
    cat("\n\n")
  } else {
    cat("Partial model from RQ2 results:")
    cat("\n\n")
    rq3.data <- heart_data[,c("HeartDisease", "ChestPainType", "Sex",
                              "ExerciseAngina")]
  }

  rq3.model.log <- create_logit_model(rq3.data)
  rq3.model.lda <- create_lda_model(rq3.data)
  rq3.model.knn <- create_knn_model(rq3.data,
                                    names(rq3.data[names(rq3.data) !=
                                                     "HeartDisease"]),
                                    "HeartDisease",
                                    k = 12)

  rq3.predicted.log <- ifelse(predict(rq3.model.log[[1]], rq3.model.log[[3]],
                                      type = "response") > .5, 1, 0)
```

```
  rq3.confusion.log <- table(Observed = rq3.model.log[[3]]$HeartDisease,
                             Predicted = rq3.predicted.log)
  rownames(rq3.confusion.log) <- c("Observed Heart Disease",
                                   "Observed No Heart Disease")
  colnames(rq3.confusion.log) <- c("Predicted Heart Disease",
                                   "Predicted No Heart Disease")
  cat("Logistic Regression: ")
  cat(paste("Misclassification Error Rate = ",
            round(1 - sum(diag(rq3.confusion.log))/sum(rq3.confusion.log), 4)))
  cat("\n")
  print(kable(rq3.confusion.log))
  cat("\n\n")

  rq3.predicted.lda <- predict(rq3.model.lda[[1]], rq3.model.lda[[3]])
  rq3.confusion.lda <- table(Observed = rq3.model.lda[[3]]$HeartDisease,
                             Predicted = rq3.predicted.lda$class)
  rownames(rq3.confusion.lda) <- c("Observed Heart Disease",
                                   "Observed No Heart Disease")
  colnames(rq3.confusion.lda) <- c("Predicted Heart Disease",
                                   "Predicted No Heart Disease")
  cat("Linear Discriminant Analysis: ")
  cat(paste("Misclassification Error Rate = ",
            round(1 - sum(diag(rq3.confusion.lda))/sum(rq3.confusion.lda), 4)))
  cat("\n")
  print(kable(rq3.confusion.lda))
  cat("\n\n")

  rq3.confusion.knn <- table(Observed = rq3.model.knn[[3]]$HeartDisease,
                             Predicted = rq3.model.knn[[1]])
  rownames(rq3.confusion.knn) <- c("Observed Heart Disease",
                                   "Observed No Heart Disease")
  colnames(rq3.confusion.knn) <- c("Predicted Heart Disease",
                                   "Predicted No Heart Disease")
  cat("k-Nearest Neighbors: ")
  cat(paste("Misclassification Error Rate = ",
            round(1 - sum(diag(rq3.confusion.knn))/sum(rq3.confusion.knn), 4)))
  cat("\n")
   print(kable(rq3.confusion.knn))
  cat("\n\n")
}
# ----- Research Question IV -----
# Subset data with those who have atypical angina
rq4_data <- heart_data[heart_data$ChestPainType == "ATA",]
rq4_data$HD <- ifelse(rq4_data$HeartDisease == "1", "heart disease present",
                      "no heart disease")
rq4.logreg <- glm(HeartDisease ~ RestingBP + Cholesterol + MaxHR, data = rq4_data,
                  family = "binomial") #create logistic regression model
# Performance of the model:
# train_data <- rq4_data[-c(1:50),] # Remove the first 50 observations
# test_data <- rq4_data[c(1:50),] # Assign the first 50 observations to test_data

n <- length(rq4_data$HeartDisease)
# Remove 50 randomly selected observations
set.seed(1)
test_obs <- sample(1:n, size = 50, replace = FALSE)
train_data <- rq4_data[-c(test_obs),]
# Assign the 50 randomly selected observations to test_data
test_data <- rq4_data[c(test_obs),]
```

```
rq4.logreg.test <- glm(HeartDisease ~ RestingBP + Cholesterol + MaxHR,
                       data = train_data, family = "binomial")
summary(rq4.logreg.test)
predicted <- ifelse(predict(rq4.logreg.test, test_data, type = "response")<0.5,"0","1")
predicted
confusion <- table(Observed = test_data$HeartDisease, Predicted = predicted)
rownames(confusion) <- c("Observed Heart Disease", "Observed No Heart Disease")
colnames(confusion) <- c("Predicted Heart Disease", "Predicted No Heart Disease")
print(kable(confusion))
plot1 <- ggplot(data = train_data) +
  geom_point(mapping = aes(x = RestingBP,
                           y = MaxHR,
                           color = HD))

plot2 <- ggplot(data = train_data) +
  geom_point(mapping = aes(x = Cholesterol,
                           y = MaxHR,
                           color = HD))

plot3 <- ggplot(data = train_data) +
  geom_point(mapping = aes(x = Cholesterol,
                           y = RestingBP,
                           color = HD))

grid.arrange(plot1, plot2, plot3, nrow = 3)
# Accuracy
sum(diag(confusion))/sum(confusion)
# Type I error
sum(confusion[2,1])/sum(confusion[,1])
# Type II error
sum(confusion[1,2])/sum(confusion[,2])
# True positive rate
sum(confusion[2,2])/sum(confusion[,2])
# True negative rate
sum(confusion[1,1])/sum(confusion[,1])
qual_pred1 <- ggplot(data = heart_data, aes(Age)) +
  geom_bar() +
  ggtitle("Distribution of age")

qual_pred2 <- ggplot(data = heart_data, aes(RestingBP)) +
  geom_bar() +
  ggtitle("Distribution of resting blood pressure")

qual_pred3 <- ggplot(data = heart_data, aes(Cholesterol)) +
  geom_bar() +
  ggtitle("Distribution of serum cholesterol")

qual_pred4 <- ggplot(data = heart_data, aes(MaxHR)) +
  geom_bar() +
  ggtitle("Distribution of maximum heart rate")

qual_pred5 <- ggplot(data = heart_data, aes(Oldpeak)) +
  geom_bar() +
  ggtitle("Distribution of oldpeak")

grid.arrange(qual_pred1, qual_pred2, qual_pred3, qual_pred4, qual_pred5, nrow = 3, ncol = 2)
```