# Heart Disease Prediction Using Logistic Regression

Olivia Zhang(V00885977), Harshit Handa(V00919882), Rory Davies(V00875731)

## The Problem

Heart disease is the term used to describe a large range of conditions that affect the structure or function of the heart. It includes cardiovascular problems such as: arrhythmia, atherosclerosis, cardiomyopathy, congenital heart defects, coronary artery disease, and heart infections (Donovan, R., 2020). According to the world health organization, heart disease is the leading cause of death worldwide with an estimated 9 million occurring in 2019 alone ("WHO reveals leading causes of death and disability worldwide: 2000-2019", 2020). Many different factors can cause heart disease, some controllable and some non-controllable. The main behavioural risk factors are a poor diet, lack of physical activity, as well as tobacco and alcohol use. These risk factors and other confounding factors will often lead to raised blood pressure, blood glucose level, and blood lipids level, which might be the implication of risk factors for the development of heart diseases.

Being able to detect early signs of heart disease can be the difference between life and death for many patients. By identifying the early signs of heart disease, at-risk patients can have time to make changes in decision making and lifestyle choices. As many heart conditions end up as chronic conditions, people's quality of life can be improved significantly with possible prevention and early detection. The overall population health can benefit from identifying the risk factors of such prevalent diseases; health authorities can use the information and implications to educate residents about the risk factors and to promote healthy lifestyles.

## Framingham Dataset

The dataset used is from a cohort study and its subsequent omni and offspring cohort studies on the residents of Framingham, Massachusetts under the direction of the National Heart, Lung, and Blood Institute (NHLBI) ("About the Framingham Heart Study", n.d.). The classification goal is to predict whether or not a patient is at a 10 year risk for heart disease. The data set included 4240 records and 15 predictor variables; each predictor represents a potential risk factor for heart disease. The following are the variables found in the dataset:

Target (Dependent) Variable:

- TenYearCHD - 10 year risk of coronary heart disease (0: Not At Risk, 1: At Risk)

Binary/Categorical Predictor Variables:

- male - gender (0: Female, 1: Male)
- education - patient's level of education (1: Some High School, 2: High School or GED, 3: Some College or Vocational School, 4: College)
- currentSmoker - whether or not the patient currently smokes (0: No, 1: Yes)
- BPMeds - whether or not patient is on blood pressure medication (0: No, 1: Yes)
- PrevalentStroke - whether or not the patient has previously had a stroke (0: No, 1: Yes)
- prevalentHyp - whether or not the patient was hypertensive (0: No, 1: Yes)
- diabetes - whether or not the patient has diabetes (0: No, 1: Yes)

Continuous Predictor Variables

- age - age of the patient
- cigsPerDay - Estimated average number of cigarettes smoked per day
- totChol - total cholesterol level
- sysBP - systolic blood pressure
- diaBP - diastolic blood pressure

- BMI - body mass index
- heart rate - heart rate (BPM)
- glucose - blood glucose level

## Exploratory Data Analysis

To examine the preliminary data, the team generated a statistical summary in R using the csv file provided (code section 2.1). The mean and median of all continuous variables are close, which indicates there are no extreme outliers from the datasets. No variables have more than 10% of missing data which ensures a high data integrity. The frequency of each binary variables' answer can be known combining the statistical summary with the tables of those binary variables produced by R (code section 2.2). There are 4240 sets of observations; 1820 of them are male participants, 2420 are female; 2095 of them are current smokers; 25 have had a stroke; 1317 of them are hypertensive; 109 are diabetic. The most of participants (1720) have some high school education out of 4135 available answers; 473 of them have college level education. There are 53 "n/a", 124 "yes," and 4063 "no" for the variable of whether the participants take hypertension medication. The researchers determined that 644 of them are at risk of developing coronary heart diseases in the next 10 years.

The correlation heatmap (figure 2) visualizes some potential issues with the dataset. First, there are some variables that are highly correlated to each other ( ie. currentSmoker/cigsPerDay, diabetes/glucose, and prevalentHyp/diaBPdiabetes) which indicates that there is collinearity present in the data. Second, the variables relating to smoking and education are not correlated to the target variable which could mean that they are insignificant predictors or they are confounding factors.

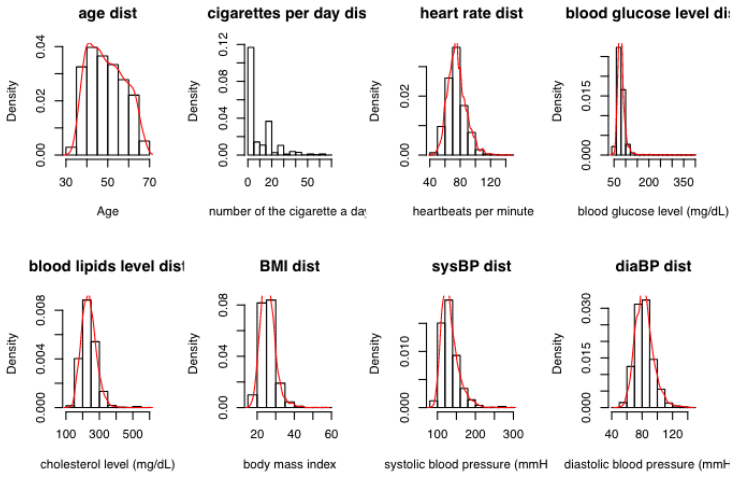


Figure 1: Histograms that show the distribution of continuous variables (from section 2.3 of code)

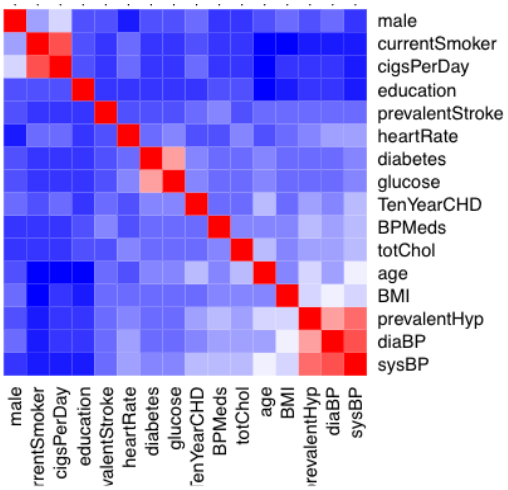


Figure 2: Correlation heat map representing correlation ranging from 0 (blue) to 1 (red)(from section 2.5 of code)

## Logistic Regression to Test Significance of the Predictors

To get the logistic regression of the data we will use the glm function in R(Appendix D) then we will check their P-values to see if the variable is significant or insignificant. After running different models we found out the best model for this data. We have printed the summary of the model below.

glm(formula = TenYearCHD ~male + age + cigsPerDay + prevalentStroke + prevalentHyp + totChol + sysBP + glucose, family = binomial(link = "logit"), data = heartDisease)

Table 1: Summary statistics of the logistic regression (computed in section 3.1 of code)

| **Coefficients:** | **Estimate** | **Std. Error** | **z value** | **Pr(>\|z\|)** |
|---|---|---|---|---|
| **(Intercept)** | -2.43889 | 0.08690 | -28.065 | < 2e-16 *** |
| **male** | 0.47442 | 0.08991 | 5.276 | 1.32e-07 *** |
| **age** | 0.066012 | 0.006269 | 10.531 | < 2e-16 |
| **cigsPerDay** | 0.020301 | 0.004075 | 4.982 | 6.29e-07 |
| **prevalentStroke** | 1.17397 | 0.41862 | 2.804 | 0.00504 ** |
| **prevalentHyp** | 0.96135 | 0.08908 | 10.793 | < 2e-16 *** |
| **totChol** | 0.002473 | 0.002473 | 2.334 | 0.0196 |
| **sysBP** | 0.013641 | 0.002791 | 4.888 | 1.02e-06 |
| **glucose** | 0.007649 | 0.001643 | 4.655 | 3.24e-06 |

**AIC**: 2918.3

Number of Fisher Scoring iterations: 5

As we can see from the summary of the Logistics Regression, all the variables are significant because their P-value is less than 0.05, AIC = 2918.3. The odds of getting a coronary heart disease for those who had a stroke before were 3.2 times the odds of getting a coronary heart disease who didn't had a stroke before and the odds of getting a coronary heart disease for those who were hypertensive were 2.6 times than those who were not.

## Model Selection

Since some of the variables in this dataset are highly correlated and/or not significant predictors of the target variable, we must use a form of feature selection to determine a model that minimizes complexity and multicollinearity. To do so, we decided to use 10-fold cross validation to fit eleven logistic elastic net models, each with a different type of penalty ($\boldsymbol{\alpha}$ = 0.0, 0.1, 0.2, …, 0.8, 0.9, 1.0). After we generated the fitted models, we then went on to predict the 10 year risk of heart disease using the testing set. Note: For our predictions we set lambda (model shrinking parameter) to be the largest it could be within 1 standard error of the minimum error (lambda.1se) so that the estimates would be more parsimonious.

To compare the performance of the eleven models, we used log loss (cross-entropy loss) instead of MSE since we were dealing with logistic regression. All models were found to have very similar values of log loss but the model where $\boldsymbol{\alpha}$ = 0.0 (ridge) was slightly less than the rest with a log loss of 0.3738263, meaning it is the best fit for the data. Due to it being a ridge model, none of the coefficients were set to zero but many were shrunk to reduce the effects of multicollinearity. Interpreting some of the largest coefficients gives us these results:

- male ($\boldsymbol{\beta}$ = 0.1917, odds ratio of 1.2113)
    - Males 1.2113 times more likely to be at risk

- BPMeds (**β** = 0.2181, odds ratio of 1.2437)
    - Patients on blood pressure medication are 1.2113 times more likely to be at risk
- prevalentStroke (**β** = 0.4226, odds ratio of 1.5259)
    - Patients who have had a stroke are 1.5259 times more likely to be at risk
- diabetes (**β** = 0.2728, odds ratio of 1.3136)
    - Patients with diabetes are 1.3136 times more likely to be at risk

Intuitively these results make sense as one would expect blood pressure medication, history of strokes, and diabetes to be signs that someone is at risk for heart disease.
Further analyzing the predictions made by the ridge model allowed us to create a confusion matrix (Table 2). The model predicted ten year heart disease risk with: 84.56% accuracy, 99.84% specificity, and 7.44% sensitivity. Accuracy represents the rate of correct predictions, specificity represents the rate of correct negative predictions (no risk of heart disease), and sensitivity represents the rate of correct positive predictions (risk of heart disease). Though the model fits the data well fairly well, it is useless for the purpose of predicting heart disease due to the extremely low sensitivity; it is only good for predicting when someone isn't at risk for heart disease. This is also made evident by the 15.30% type II error rate.

**Table 2**: Confusion matrix and performance measures for prediction. (Calculations from section 4.4 in code)

| TenYearCHD Training | Actual | | | Accuracy = 0.8456<br>Sensitivity = 0.0744<br>Specificity = 0.9984<br>Type I Error Rate = 0.0014<br>Type II Error Rate = 0.1530 |
|---|---|---|---|---|
| Predicted | | Not At-Risk | At-Risk | |
| | Not At-Risk | 610 | 112 | |
| | At-Risk | 1 | 9 | |

Note: We used a probability threshold of 0.5 to classify the predictions.

## Conclusion

The logistic regression we used to try and predict ten year heart disease risk was not sufficient for our goal due to the low sensitivity of the model. Comparing our result with other Kaggle projects on this topic showed that this was the case for nearly all logistic predictive  models. It would be more useful to try and use a different method of prediction for this dataset.

## References


Kaggle Competition:
https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression?fbclid=IwAR0DJTjgHzbBPi4Kv8sdy3aMXn_5qOnTpIAjh-D-kTmJdyu6jpwaaTQWhbM

About the Framingham Heart Study. (n.d.). https://framinghamheartstudy.org/fhs-about/

Cardiovascular Diseases. (2017, May 17). https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Donovan, R. (2020, February 27). *Everything You Need to Know About Heart Disease.* https://www.healthline.com/health/heart-disease#risk-factors

WHO reveals leading causes of death and disability worldwide: 2000-2019. (2020, December 9). https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019