**Logistic Regression Analysis of Myocardial Infarction Incidence**

Evan Donald

April 27, 2019

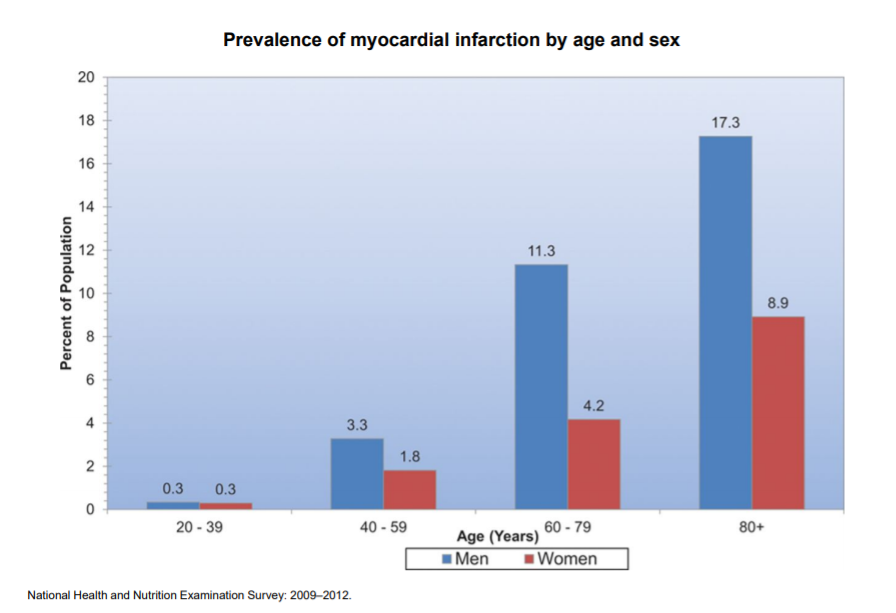
**Abstract**

Logistic regression is one of the popular ways to model binary data. In the paper we use it to model myocardial infarction, cardiovascular disease incidence (0 = myocardial infarction occurred, 1 = no myocardial infarction occurred). The data used is from NHANES III. Personal background information (e.g., age, race, sex), diet (e.g., frequency of cheese, frequency of processed meats), smoking (e.g., interviewed individual smoker, member of household who smokes, number of members in household who smoke), and health (e.g., congestive heart failure, high blood pressure, high cholesterol, chest pain, difficulty walking) were used as predictor variables. The results indicate that sex, age, congestive heart failure, chest pain, and frequency of cheese consumption are significant in predicting whether an individual has had a heart attack. The estimated equation is useful in predicting whether an individual is prone to the risk of heart attack based on his/her personal information, diet, smoking habits, and health.

**Introduction**

Heart disease is the number one cause of death in the United States (Heart Disease Facts, n.d.). A heart attack is one of the various forms of heart disease, including arrhythmia, coronary artery disease, heart failure and many others. A heart attack, or a myocardial infarction, occurs when blood flow to the heart is blocked. The blockage is often a buildup of fat, cholesterol, and other substances, which form plaque in the coronary arteries. Once the plaque breaks away and forms a clot the blood flow is stopped, and the heart muscle could be damaged or destroyed (Heart attack, 2018). According to Centers for Disease Control and Prevention someone in the United States has a heart attack every forty seconds. Every year around 790,000 Americans have a heart attack (Heart Attack Facts, n.d.). Of those 790,000, almost 580,000 are first time heart attacks and 210,000 happen to people who have previously had a heart attack (Heart Attack Facts, n.d.).

Studies show that the number of heart disease deaths vary based on personal information, for example, race, sex, and age. According to KFF (Kaiser Family Foundation), heart disease is the cause of 164.4 deaths per 100,000 population in whites, and the cause of 202.4 deaths per 100,000 population in African Americans (Number of Heart Disease, 2019). For other populations heart disease causes 90.1 deaths per 100,000 population (Number of Heart Disease, 2019). Sex also varies the number of heart disease deaths. According to KFF, for every 100,000 males, there are 209.0 deaths due to heart disease, and for every 100,000 females, there are 129.6 deaths due to heart disease (Number of Heart Disease, 2019). According to the American Heart Association, the percentage of the United States population that have had a heart attack is shown in the chart below.

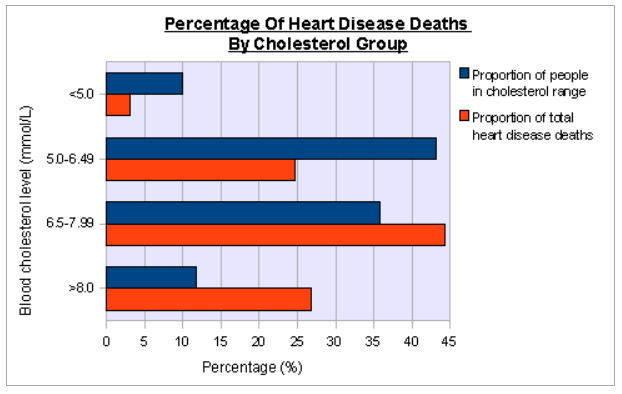


Regardless of gender, as the age increases, the percentage, of individuals who have had a heart attack, increases (Prevalence, 2015).

Studies also show that a person’s diet can increase, or decrease, the risk of heart disease. A poor diet can drastically increase the chance of a person having a heart attack. BetterHealth states that a person should avoid eating processed foods, and switch to low-fat or non-fat dairy products (Heart disease and food, 2012).

Smoking has proven to cause around one-fifth of deaths due to heart disease in the United States (Smoking, 2013). A smoker is between twice and four times more likely to have a heart disease (Smoking, 2013). Also, a smoker has around 70% higher death rate from heart disease (Smoking, 2013).

Other health concerns can also increase risk of heart disease. High cholesterol can increase one’s risk at heart disease drastically. As you can see in the chart below, a person with a cholesterol above 8mmol/L is more than five times as likely to develop a heart disease than a person with a level below 5mmol/L (Wilson, 2008).



According to the CDC high blood pressure also puts a person at risk of heart disease (Heart Disease Facts, n.d.).

**Data Description:**

In this paper we used data from the Third National Health and Nutrition Examination Survey (NHANES III). The survey was conducted between the years 1988 and 1994 and contains data for 33,994 people. The National Center for Health Statistics of the Centers for Disease Control and Prevention collects, analyzes, and disseminates data on the health status of United States residents.

The data is stored in three files; the Household Adult file, the Laboratory File, or the Examination file. During the logistic regression procedures, the number of samples was reduced because of the removal of two answers to the criterion variable. All the samples with the answers “Blank but applicable” and “Don’t know” were removed. Missing data also eliminated samples, leaving the final data set.

The selection of variables was determined based on prior research into the causes of heart attacks. The investigation has thirteen selected variables separated into four categories; personal information, smoking habits, diet, and health.

**Exploratory Analysis of Selected Variables:**

*Personal Information:* These variables include race, sex, and age.

The race variable is separated into four different races, White (1), African American (2), and Other (3). White makes up 75.04%, totaling 2633 people. African American makes up 22.94% and has 805 people. Finally, the Other category makes up the remaining 2.02% and totals 71 people.

Sex is separated into male (1) and female (2). There are more males than females in the data set.

Age is measured in months at the time the question was asked. The mean of the age variable, as seen below, is 725.887 months. This can be converted to 60.491 years.

*Smoking:* There are three variables related to smoking used in the investigation.

The first variable related to smoking is current smoker. If the interviewed individual is a smoker the value is 1, and if not, the value is 2. People who smoke make up 35.71% of the data, totaling 1253 people. People who do not smoke make up the remaining 2256 people.

The other two variables in the smoking category relate to members of the household smoking. The first variable is if the person interviewed lives with a smoker. The following bar graph shows the data. If the value is 1, there is a smoker in the house. The second variable is the number of members in the household who smoke. Most people do not live with a smoker. Of those who do live with a smoker, the most common number of smokers is one.

*Diet:* The two variables used for diet are the consumption of cheese, and the consumption of processed meat.

The mean for times cheese is eaten per month is 9.892. The mean for number of times processed meats are eaten per month is 7.539.

*Health:* Five variables were used that have been categorized as health variables. The variables are congestive heart failure, high blood pressure, high cholesterol, chest pain, and difficulty walking. The frequencies of the variables are listed.

Congestive heart failure: 249 people experienced congestive heart failure.

High blood pressure: 1449 people have high blood pressure.

High cholesterol: 1245 people have high cholesterol.

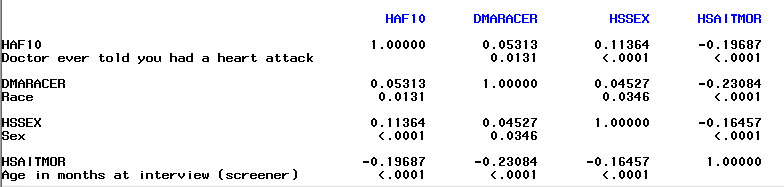
Chest pain: 1234 people experienced chest pain.

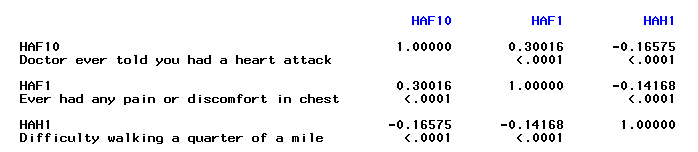
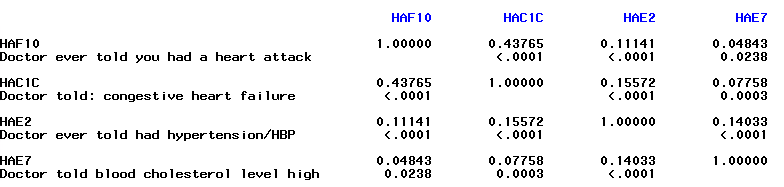
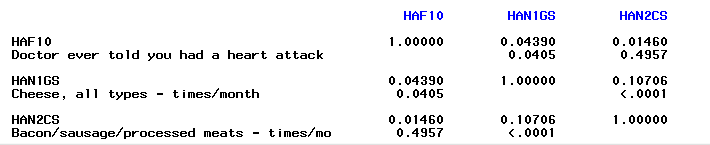
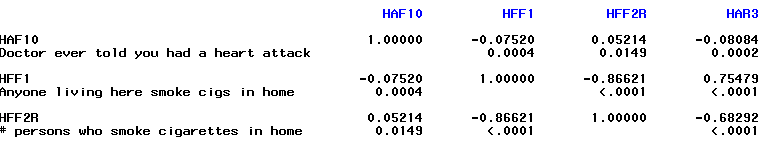
Difficulty walking: 2637 people have no difficulty walking, 428 people have some difficulty walking, 169 have much difficulty walking, and 275 are unable to walk.

*Dependent Variable:*  The dependent variable for the investigation was the yes or no answer to the question “Has a doctor ever told you that you had a heart attack?”. In the sample, 10.60% of people have had a heart attack before. Of the 3509 people in the data set, 372 have experienced a heart attack.

**Bivariate Relation**

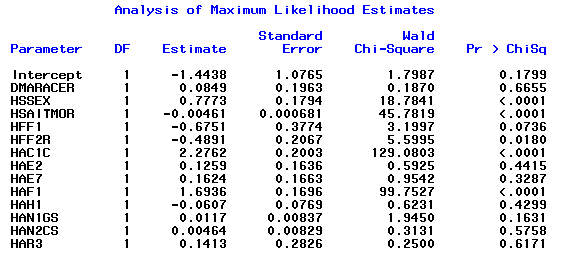
The correlations of the variables in relation to heart attack are seen below.





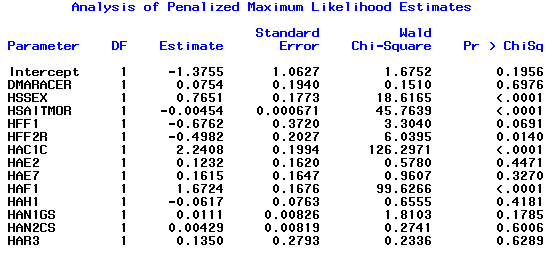
**Logistic Regression Analysis**

Since the dependent variable can only be one of two outcomes (1 or 2), logistic regression is used over linear regression or multiple linear regression. Logistic regression created a model that allows the predictor variables to predict the outcome of the dependent variable. The dataset was split into two parts. The first part of the data, called the training set, contained 2,179 (or 60%) of the samples. The second data set, called the validation set, had the remaining 40% of the data. The training set was used to create a model to predict the dependent variable. The validation data is used to assess the fitted model. The model can be seen below.



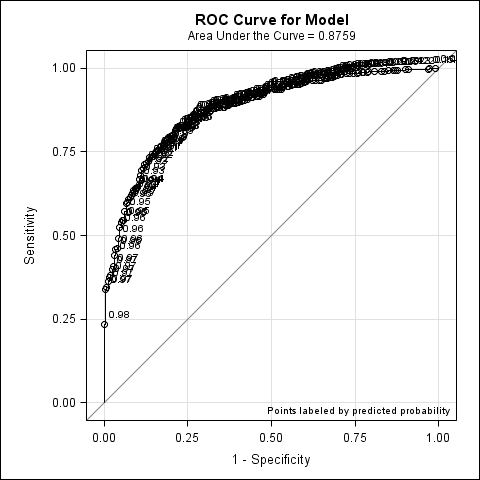
Using a p-value of 0.05, the variables sex, age, number of smokers in household, congestive heart failure, and chest pain are significant. Since the model was created descending, if a variable has a positive coefficient, the likelihood of a heart attack increases. Also, if the coefficient is negative, the chance of heart attack decreases.

Since the likelihood of a person in the data set who had a heart attack is around 10%, a logistic regression for rare events was also used. The Firth logistic regression is shown below.



**Results**

The logistic regression produced a formula to predict the probability of a person having a heart attack in the past. The formula was used to predict the probability of heart attack for each person in the validation set. If the probability was above 85%, the person was marked as having a previous heart attack, and if the probability was less than 85%, they were marked as not having a previous heart attack. Of the 1,330 people in the validation set, the model correctly predicted 1,181. That results in the model being 88.80% accurate. The ROC curve can be seen below.



Since the ROC curve follows the left-hand border and the top border, and is away from the 45-degree diagonal, the test is accurate.

Since the event of a person having a previous heart attack was around 10%, the event was not rare enough for the Firth logistic regression to have any effect. The Firth logistic regression model predicted the probability of a previous heart attack at the same accuracy as the logistic regression model.

**Discussion**

Logistic regression was used to formulate a model to predict whether a person has had a previous heart attack. Of the thirteen independent variables, only five were significantly related to previous heart attacks. The variables that were significant were sex, age, number of smokers in the household, congestive heart failure, and chest pain. The three most significant variables are sex, age, and chest pain. The model was more accurate predicting the probability that a person did not have a heart attack. With a 70.80% accuracy predicting previous heart attack, the model is accurate.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Sample | Model | Model Accuracy |
| Heart Attack | 113 | 80 | 70.80% |
| No Heart Attack | 1217 | 1101 | 90.47% |

The overall model could have been more accurate if the deciding probability was 50% instead of 85%. This brought the overall accuracy to 92% but dropped the accuracy of a previous heart attack to 27%.

**Summary**

The model showed that of the thirteen independent variables, five are significant in predicting heart attacks. It has been shown that males are more likely to have had a previous heart attack. Also, as age increases, the probability of a previous heart attack increases. Also, as the number of smokers in the person household increases, the chance of heart attack increases as well. People who have congestive heart failure and people who experience chest pain are also more likely to have had a heart attack. This model can be used to predict the likelihood of a person having a heart attack in the future and predict the probability that a person had a previous heart attack and did not know it.

**References**

Heart attack. (2018, May 30). Retrieved April 15, 2019, from <https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106>

Heart Attack Facts & Statistics. (n.d.). Retrieved April 15, 2019, from <https://www.cdc.gov/heartdisease/heart_attack.htm>

Heart disease and food. (2012, September 30). Retrieved April 15, 2019, from <https://www.betterhealth.vic.gov.au/health/ConditionsAndTreatments/heart-disease-and-food>

Heart Disease Facts & Statistics. (n.d.). Retrieved April 15, 2019, from <https://www.cdc.gov/heartdisease/facts.htm>

Number of Heart Disease Deaths per 100,000 Population by Race/Ethnicity. (2019, February 11). Retrieved April 15, 2019.

Prevalence of coronary heart disease by age and sex. (2015). Retrieved April 15, 2019, from <https://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_449846.pdf>

Smoking & Disease. (2013, December). Retrieved April 15, 2019, from <https://my.clevelandclinic.org/health/articles/17488-smoking>

Wilson, A. (2008, January 25). Cholesterol And Coronary Heart Disease – The Facts. Retrieved April 15, 2019, from <http://healthhubs.net/heartdisease/cholesterol-and-coronary-heart-disease-the-facts/>