James Taylor

Assignment 2

University of Maryland University College

DATA630-9041 Spring 2020

Dr. Gates

james.taylor.ps3@gmail.com

**Introduction**

Cardiovascular disease is the leading cause of death in the United States, accounting for 25% of all deaths each year (“Coronary Heart Disease”, n.d.). Heart attacks are a dramatic consequence of it, and recently about 805,000 Americans have a heart attack each year (Fryar CD, Chen).

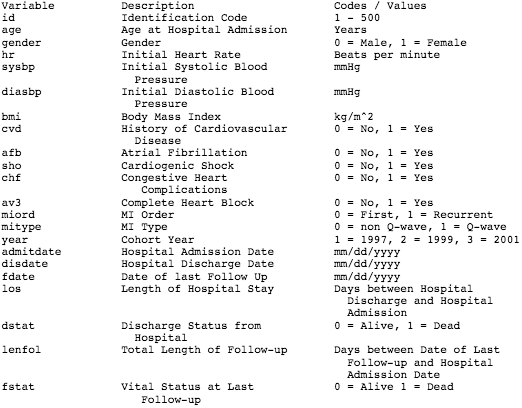
Heart attacks used to be deadlier. It was only a few decades ago where a heart attack would kill half of the patients in a few days (“How Heart Attacks became Less Deadly”, n.d.). Since 1950, death rates from cardiovascular disease (CVD) have declined 60% (“Decline in Deaths from Heart Disease and Stroke -- United States, 1900-1999”, n.d.). Researchers learned more about the causes and treatments for it, like aspirin and blood-thinning drugs.

Heart attacks are often survivable, but people still die. This paper will train a model that will look at the likelihood of death following release from hospital. Inspecting the model’s ability to accurately predict (on testing data) which patients will live and die, and which variables are most significant in driving that accuracy, will help physicians understand the factors that drive a patient to die after discharge following a heart attack. If the model is inaccurate, it will be evidence that heart attacks are complicated, intricate events that are significantly different between individuals. While the past century has brought gains to heart attack survivorship, the goal of this paper is to determine what factors are associated with a patient dying today, and how well those can be used to predict an individual patient’s chances using a logistic regression.

**Analysis and Model Development**

About Dataset

The source of this data was Dr. Robert J. Goldberg of the Department of Cardiology at the University of Massachusetts Medical School. It follows patients who were admitted into the hospital for a heart attack in 1997, 1999 and 2001. Either the patients died at the hospital, or they were discharged and checked up on some later time. The below figure lists the variable names, the descriptions and the values and meanings.



*Source: Hosmer, D.W. and Lemeshow, S. and May, S. (2008) Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition, John Wiley and Sons Inc., New York, NY*

The dataset has 500 observations on 22 variables. There is demographic data on the patient like gender, age, body mass index and number of myocardial infarction (commonly called a heart attack). As the patient is admitted, their heart-rate and blood pressure are taken down. There is some categorical information on what kind of heart attack occurred and what is seen on devices tracking the heart. Dates are included for the admittance, discharge and follow-up on each patient. The length of follow-up and stay at hospital use those day to determine how many days passed.

Data Processing

The dataset had no missing variables or obvious outliers. There were variables that were unnecessary, which were removed. One was ID Number, which is an identification code that identifies the observation. The others were the dates of admittance, discharge, follow-up and year. The exact dates and year do not matter, but the number of day between them in each observation does. Our model is meant to find relationships independent from time, so the year variable was excluded.

Observations where the patient died while in the hospital were removed, along with the variable indicating it (‘dstat’). This was because it would have biased the model. These observations still had a LENFOL value even when the patient died in the hospital. Those LENFOL values are meaningless to our model, because the purpose is to find variables that drive mortality once a patient is discharged. There were 39 such cases in the dataset. The remaining variables were all numerical, many of where Boolean. After all the removal of unneeded variables/observations the dataset had 461 observations on 16 variables.

Exploratory Data Analysis

 Age is likely to be a contributing factor on surviving a heart attack. Seen in the lower figure, the older a patient is, the less their body may be able to handle the stresses on their body as a result of the heart attack. Once that patient is discharged from a hospital, older people are also more likely to die from causes other than the heart-attack or its complications, like cancer or illness.

Gender appears to play a role in when someone is likely to have a heart attack, as the figure below shows. The age distribution of male heart attack victims have about the same frequency from 50 years old until 80. More interestingly, however, is women tend to have heart attacks later in life; most common about age 85. Because the model will control for both age and gender, this relationship won’t bias the estimates.

 The distribution of days has a minimum value of six days and a max of six and a half years; a wide range. There appears to be common ‘check-ins’ that a patient is followed-up on at 500, 1200 and 2000 days later. The proportion of deceased patients at follow-ups is not uniform across the distribution. In this dataset, every patient that was followed-up before about 350 days were all deceased. The variable is the time of the last follow-up, so if a check-in revealed a patient was dead, there would be no reason to have another follow-up. It appears discharged patients are at most risk soon after leaving the hospital.

The relationship between obesity and heart attacks is intuitive, poor diet and exercise puts stress on the cardiovascular system. Age would almost certainly be a factor as well, since the longer someone is obese the more damage is done on their organs. The figure to the left shows both those relationships. There proportion of deaths increases as age increases, the blue dots between age 60-80 are significantly outnumber the ones between 40-60. The points appears to have a downward trend, so that as someone’s age increases, their BMI decreases. It’s possible this is because obese people don’t die in old age, they die younger.

Model

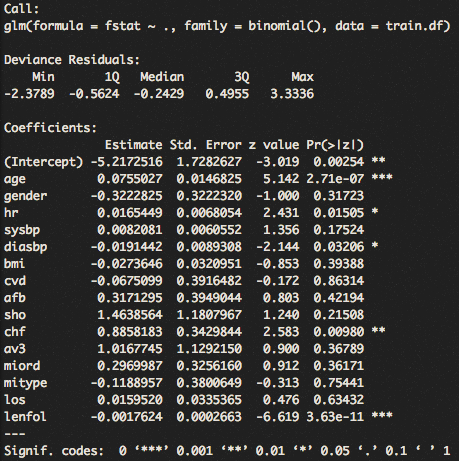
The method used to predict a patient’s likelihood of death following discharge was a logistic regression. It is similar to a linear regression where a mathematical formula attempts to find a linear equation that best fits the data. The logistic regression depends on the dependent variable being Bernoulli. Our model’s dependent variable will be the follow-up state of the patient (alive/dead.

The model will estimate the relationship between each variable and its relationship to the dependent variable. Just like a linear regression, coefficients are estimated for the independent variables. They are interpreted to be the effect on the probability as the independent variable in increased by one, all else equal. The coefficients are estimated in log of odds, but easily transformable into percentages. It is a monotonic transformation, so positive coefficients mean increased chances of the dependent being true, and false for negative.

The dependent variable will be the follow-up status of the patient (alive/dead = 0/1). The independent variables were Age, Gender, Initial Heart Rate, Initial Systolic Blood, Initial Systolic Blood Pressure, Initial Diastolic Blood Pressure, Body Mass Index, History of Cardiovascular Disease, Atrial Fibrillation, Cardiogenic Shock, Congestive Heart Complications, Complete Heart Block, MI Order, MI Type, Length of Hospital Stay, and Total Length of Follow-up in. There was a total of 15 independent variables. The computation was run using the glm function in R.

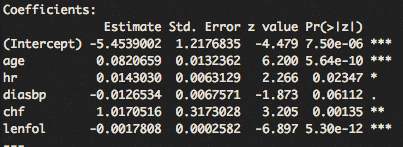
To test the accuracy of the model, the dataset was randomly split into training and test data. The training sample was to be 85% of the whole dataset, which would be 388 observations, leaving 73 observations to test the model’s accuracy (test data).

**Results and Model Evaluation**

****The model’s estimated coefficients on the training data are in the right figure. The level of significance the model places on a coefficient is the statistical confidence of the influence that variable has on the dependent. The summary of the model shows five variables that are statistically significant at the 1% level. Those are age, heart rate, diastolic blood pressure, congestive heart complications and length of follow-up.

The model’s output is the log-odds of the dependent variable being one, which can be transformed into percentages. These percentages were rounded to a zero or one so that it can be compared to the true value in the training data. The rounding threshold was set at .5, so that any percentage there or above was converted to one, and below a zero. The model outputs were 85% accurate on the training data (total correct / total predictions on 388 observations).

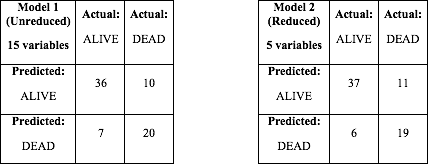
Since the model had so many statistically insignificant coefficients, the model could be simplified without sacrificing much accuracy. Using the step function to remove an insignificant variable until all variables left are significant, this reduced model had only five variables. They were age, heart rate, diastolic blood pressure, congestive heart complications and length of follow-up, which were the same five that were significant in the original model. This reduced model was slightly more accurate on the training data, correctly predicting 336 of the 388 observations (86.5%).

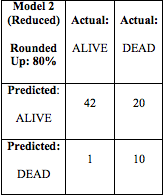
The two most significant variables are age and length of follow-up. As a heart attack patients age increases, their likelihood of being dead at checkup increases. The further a patient’s follow-up is in the future, the less likely they are to be dead at follow-up. One theory to explain this is obvious, which is surviving patients will always have the oldest follow-up length. A patient that dies will stop getting follow-ups, but surviving patents can get follow-ups long into the future. Another theory involves doctors knowing the outlook for patients with severe heart attacks isn’t good, so they get checked-up on more recently along with die sooner. Discharged patients are at most risk right after discharge.

The other variables add lots of predictive validity to the model. Congestive heart complications are particularly bad for the patient’s survival if they have it, given the high value of the estimate (1.01). The higher the patient’s heart rate is the worse their chances of survival are. The lest statistically significant variable was diastolic blood pressure, which was negatively correlated with a patient’s chances of survival.

Model Accuracy on Test Data

The models had the same accuracy on the test data. Below are the confusion matrices for both. The potential errors this model can make is a false positive or false negative, or Type I and II errors respectively. In these models, a true positive would be accurately predicting a patient to die. A Type I error would be predicting a patient will die, but they end-up living. Predicting a patient will live, but in fact they die is a Type II error.



The confusion matrices above show both models correctly predicted 56 of the 73 observations in the test dataset. Because the output of the model is a percentage chance that patient will die, the threshold of when those values are rounded to either zero or one can be adjusted. The confusion matrices above were rounded at the .5 threshold. Increasing the threshold of the model’s confidence to .8 would show how the model performed on its very confident predictions. This table (previous page) shows the confusion matrix where model 2’s predictions above 80% were rounded to predict that patient would die, proving 91% of its predictions a patient would die were correct. The model does suffer an accuracy penalty when minimizing Type I or II errors, but it accomplished the user’s purpose better.

 The ROC curve (Receiver Operator Characteristic) illustrates this tradeoff. The curve tracks the probability of a predicted positive will be positive against the probability a predicted positive will be negative. The figure on the left is the ROC curve for the reduced (5 variable) model on the test data. The above model with the rounding threshold at 10% instead of 50% demonstrates this. When the model rounds at 50%, the true positive rate is 63% and false positive rate is 14%. The model with rounding at 80% had a true positive rate of 33% and false negative at 2%. The grey dotted line shows a randomly guessing model, so the farther above that line our ROC curve is, the better our model is at predicting if a patient lives or death. The goal is to maximize the difference between the true positive rate and false positive rate.

 The residuals plot shows how the reduced model fits the training data. The blue dots in a line at the top are patients who died, the other below are ones that lived. The residual is the difference between the models out (percent likely dead at follow-up) and the patient’s actual follow-up status. The blue dots furthest from the horizontal grey line are observations where the model performed the worst. It was very confident in its prediction but was incorrect. The black line is a non-parametric local regression.

**Conclusion**

Overall, the reduced model was able to accurately predict 76% on the test data. In test cases where the model was confident (greater than 80%) a patient would die, it had a 91% of its death predictions were correct. That level of accuracy achieved by only five variables means that who will die following a heart attack is generally predictable, and not complicated events that are vary wildly between people.

The significant variables were age, heart rate, diastolic blood pressure, congestive heart complications and length of follow-up. Three of these variables (heart rate, blood pressure and complication) are observed once a patient enters the hospital, so they have already occurred before reaching medical professionals. That may be one reason why even with the advancements in care over the years, these variables are still driving deaths because the damage is occurring before they reach doctors. It may be reasonable to view the three variables simply as the severity of the heart attack, at which point it would not be surprising these are the variables that are driving mortality.

Further research should investigate how much of a patient’s likelihood of death comes from the condition they were in when they arrived at a hospital. Physicians can only do so much to help a patient’s health from the state they arrived in. Analysis on the symptoms of a high heart rate, low blood pressure and congestive heart complications and how they may be causing the increase in death rates. Learning how these symptoms can be alleviated may significantly increase survivorship because these factors are likely driving the deaths.

References

B Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010. NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Retrieved from <https://www.cdc.gov/nchs/data/databriefs/db103.pdf>

Coronary Heart Disease. (n.d.). National Heart, Lung and Blood Institute. Retrieved from <https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease>.

How Heart Attacks became Less Deadly. (n.d.). Harvard Health Publishing. Retrieved from <https://www.health.harvard.edu/healthbeat/how-heart-attacks-became-less-deadly>

Decline in Deaths from Heart Disease and Stroke -- United States, 1900-1999. (n.d.). Center For Disease Control. Retrieved from <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm4830a1.htm>