International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 11
ISSN: 2454-4248
55 – 58

Framingham Heart Study

Ashutosh Gupta
Veermata Jijabai Technological Institute (VJTI)
Mumbai, India.
ashutoshgupta2599@gmail.com

Vidhi Khathuria
Thadomal Shahani College of Engineering
Mumbai, India.
vbkhathuria@gmail.com

***Abstract*—**This paper describes the Framingham Heart Study one of the most important epidemiological studies ever conducted, and the underlying analytics that led to our current understanding of cardiovascular disease.

The logistic regression algorithm is used to analyse the Framingham data set and predict the heart risk of a patient.

Keywords- *Framingham heart study, CHD, logistic Regression, Risk Factors*

*****

I. INTRODUCTION

There were early misconceptions in the first half of the 20th century about blood pressure.
High blood pressure, dubbed essential hypertension, was considered important to force blood through arteries, and it was considered harmful to lower blood pressure.
Today, of course, we know better.

In the late 1940s, the US government set out to better understand cardiovascular disease.
The plan was to track a large cohort of initially healthy patients over their lifetimes.
A city was chosen, the city of Framingham, Massachusetts, to be the site for the study.
Framingham has an appropriate size.
It's not too large, it's not too small.
It has a stable population that doesn't move too much.
And the doctors and residents were quite cooperative.
So in 1948, the Framingham Heart Study started.

The study included 5,209 patients, aged 30 to 59.
Patients were given a questionnaire and an examination every two years.
During this examination, their physical characteristics were recorded, their behavioral characteristics, as well as test results.
Exams and questions expanded over time, but the key in the study was that the trajectory of the health of the patients was followed during their entire lifespan. We will build models using the Framingham data to predict and prevent heart disease.

II. PROPOSED APPROACH

In this paper, we'll be using analytical models to prevent heart disease.
The first step is to identify risk factors, or the independent variables, that we will use in our model.
Then, using data, we'll create a logistic regression model to predict heart disease.
Using more data, we'll validate our model to make sure it performs well out of sample and on different populations than the training set population.
Lastly, we'll discuss how medical interventions can be defined using the model.

We'll be predicting the 10-year risk of coronary heart disease or CHD.
This was the subject of an important 1998 paper introducing what is known as the Framingham Risk Score.
This is one of the most influential applications of the Framingham Heart Study data.

We'll use logistic regression to create a similar model.
CHD is a disease of the blood vessels supplying the heart.
This is one type of heart disease, which has been the leading cause of death worldwide since 1921.
In 2008, 7.3 million people died from CHD.
Even though the number of deaths due to CHD is still very high, Age-adjusted death rates have actually declined 60% since 1950.
This is in part due to earlier detection and monitoring partly because of the Framingham Heart Study.

Before building a logistic regression model, we need to identify the independent variables we want to use.
When predicting the risk of a disease, we want to identify what are known as risk factors.
These are the variables that increase the chances of developing a disease.
Identifying these risk factors is the key to successful prediction of CHD.

We'll focus on the risk factors that they collected data for in the original data collection for the Framingham Heart Study.
We'll be using an anonymized version of the original data that was collected.
This data set includes several demographic risk factors-- the sex of the patient, male or female; the age of the patient in years; the education level coded as either 1 for some high school, 2 for a high school diploma or GED,

 ISSN: 2454-4248
55 – 58

3 for some college or vocational school, and 4 for a college degree.
The data set also includes behavioral risk factors associated with smoking-- whether or not the patient is a current smoker and the number of cigarettes that the person smoked on average in one day.
Medical history risk factors were also included.
These were whether or not the patient was on blood pressure medication, whether or not the patient had previously had a stroke, whether or not the patient was hypertensive, and whether or not the patient had diabetes.

Lastly, the data set includes risk factors from the first physical examination of the patient.
The total cholesterol level, systolic blood pressure, diastolic blood pressure, Body Mass Index, or BMI, heart rate, and blood glucose level of the patient were measured.

Now that we have identified a set of risk factors, we will use this data to predict the 10 year risk of CHD.
First, we'll randomly split our patients into a training set and a testing set.
Then, we'll use logistic regression to predict whether or not a patient experienced CHD within 10 years of the first examination.
All of the risk factors were collected at the first examination of the patients.
After building our model, we'll evaluate the predictive power of the model on the test set.
We have data for 4,240 patients and 16 variables.
We have the demographic risk factors male, age, and education; the behavioral risk factors currentSmoker and cigsPerDay; the medical history risk factors BPMeds, prevalentStroke,
prevalentHyp, and diabetes; and the physical exam risk factors totChol, sysBP, diaBP, BMI, heartRate, and glucose level.
The last variable is the outcome or dependent variable, whether or not the patient developed
CHD in the next 10 years.
We split our data into a training set and a testing set
.
Here, we'll put 65% of the data in the training set.
When you have more data like we do here, you can afford to put less data in the training set and more in the testing set.
This will increase our confidence in the ability of the model to extend to new data since we have a larger test set, and still give us enough data in the training set to create our model.

Now,

Now we're ready to build our logistic regression model using the training set.
We'll use here where we predict our dependent variable using all of the other variables in the data set as independent variables.

Let's take a look at the summary of our model.

It looks like male, age, prevalent stroke, total cholesterol, systolic blood pressure, and glucose are all significant in our model.
Cigarettes per day and prevalent hypertension are almost significant.
All of the significant variables have positive coefficients, meaning that higher values in these variables contribute to a higher probability of 10-year coronary heart disease.

III. RESULTS

Now, let's use this model to make predictions on our test set.

Now, let's use a threshold value of 0.5 to create a confusion matrix.
.
With a threshold of 0.5, we predict an outcome of 1, the true column, very rarely.
This means that our model rarely predicts a 10-year CHD risk above 50%. The accuracy of this model is 1069 plus 11, divided by the total number of observations in our data set, 1069 + 6 + 187 + 11.
So the accuracy of our model is about 84.8%.
Comparing this to the accuracy of a simple baseline method.

The more frequent outcome in this case is 0, so the baseline method would always predict 0 or no CHD.
This baseline method would get an accuracy of 1069
+ 6-- this is the total number of true negative cases--divided by the total number of observations in our data set, 1069 + 6 + 187 + 11.
So the baseline model would get an accuracy of about 84.4%.
We will compute the out-of-sample AUC.
So we have an AUC of about 74% on our test set, which means that the model can differentiate between low risk patients and high risk patients pretty well.
As we saw in R, we were able to build a logistic regression model with a few interesting properties.
It rarely predicted 10-year CHD risk above 50%.

So the accuracy of the model was very close to the baseline model.
However, the model could differentiate between low risk patients and high risk patients pretty well with an out-of-sample AUC of 0.74.
Additionally, some of the significant variables suggest possible interventions to prevent CHD.
We saw that more cigarettes per day, higher cholesterol, higher systolic blood pressure, and higher glucose levels all increased risk.

IV. CONCLUSION

A compressive study of framingham heart data set has been done using logistic regression algorithm. Although the accuracy of the final result is higher than the baseline model, there is still further scope of improvement.

 ISSN: 2454-4248
55 – 58

REFERENCES

- https://en.wikipedia.org/wiki/Framingham_Heart_Study
- https://www.nih.gov/sites/default/files/about-nih/impact/framingham-heart-study.pdf
- L. R. Ott and M. Longnecker, An Introduction to Statistical Methods and Data Analysis, Brooks/Cole, Belmont, Canada, 6TH edition, 2010.
- Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological), 20(2), 215-242.

```
> summary(framingham)
      male              age        education     currentSmoker      cigsPerDay
 Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000   Min.   : 0.000
 1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 0.000
 Median :0.0000   Median :49.00   Median :2.000   Median :0.0000   Median : 0.000
 Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941   Mean   : 9.006
 3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:20.000
 Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000   Max.   :70.000
                                  NA's   :105                      NA's   :29
     BPMeds       prevalentStroke     prevalentHyp       diabetes         totChol
 Min.   :0.00000   Min.   :0.000000   Min.   :0.0000   Min.   :0.00000   Min.   :107.0
 1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:206.0
 Median :0.00000   Median :0.000000   Median :0.0000   Median :0.00000   Median :234.0
 Mean   :0.02962   Mean   :0.005896   Mean   :0.3106   Mean   :0.02571   Mean   :236.7
 3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:263.0
 Max.   :1.00000   Max.   :1.000000   Max.   :1.0000   Max.   :1.00000   Max.   :696.0
 NA's   :53                                                             NA's   :50
     sysBP            diaBP             BMI           heartRate         glucose
 Min.   : 83.5   Min.   : 48.0   Min.   :15.54   Min.   : 44.00   Min.   : 40.00
 1st Qu.:117.0   1st Qu.: 75.0   1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00
 Median :128.0   Median : 82.0   Median :25.40   Median : 75.00   Median : 78.00
 Mean   :132.4   Mean   : 82.9   Mean   :25.80   Mean   : 75.88   Mean   : 81.96
 3rd Qu.:144.0   3rd Qu.: 90.0   3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00
 Max.   :295.0   Max.   :142.5   Max.   :56.80   Max.   :143.00   Max.   :394.00
                                 NA's   :19      NA's   :1        NA's   :388
      TenYearCHD
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.1519
 3rd Qu.:0.0000
 Max.   :1.0000
```

```
> # Predictions on the test set
> predictTest = predict(framinghamLog, type="response", newdata=test
>
> # Confusion matrix with threshold of 0.5
> table(test$TenYearCHD, predictTest > 0.5)

    FALSE TRUE
  0  1182    5
  1   200   11
> # Accuracy
> (1182+11)/(1182+11+200+5)
[1] 0.8533619
```

ISSN: 2454-4248
55 – 58

```
> # Logistic Regression Model
> framinghamLog = glm(TenYearCHD ~ ., data = train, family=binomial)
> summary(framinghamLog)

Call:
glm(formula = TenYearCHD ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9148  -0.5972  -0.4254  -0.2845   2.8180

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.170671   0.860672  -9.493  < 2e-16 ***
male             0.513022   0.130232   3.939 8.17e-05 ***
age              0.064551   0.008098   7.971 1.57e-15 ***
education       -0.012780   0.060012  -0.213  0.83137
currentSmoker    0.050538   0.185866   0.272  0.78570
cigsPerDay       0.018152   0.007440   2.440  0.01470 *
BPMeds           0.331498   0.285651   1.161  0.24585
prevalentStroke  1.157806   0.558173   2.074  0.03805 *
prevalentHyp     0.265059   0.165560   1.601  0.10938
diabetes        -0.400224   0.404936  -0.988  0.32297
totChol          0.003204   0.001330   2.409  0.01601 *
sysBP            0.011836   0.004441   2.665  0.00769 **
diaBP           -0.004978   0.007765  -0.641  0.52143
BMI              0.008722   0.015534   0.561  0.57450
heartRate       -0.004891   0.005043  -0.970  0.33214
glucose          0.008883   0.002840   3.128  0.00176 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2174.3  on 2590  degrees of freedom
Residual deviance: 1934.7  on 2575  degrees of freedom
  (165 observations deleted due to missingness)
AIC: 1966.7

Number of Fisher Scoring iterations: 5
```