Framingham Heart Study

(Adapted from Bertsimas Chapter 7)

Introduction

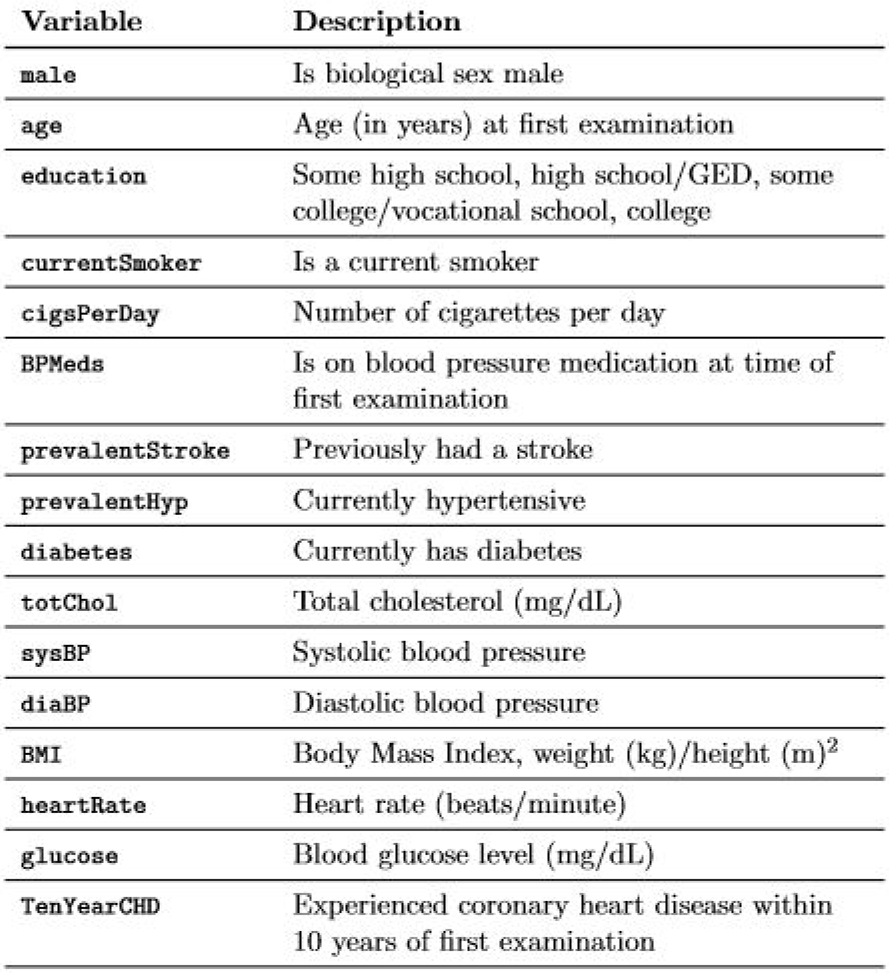
Heart disease is the leading cause of death worldwide. About 7.4 million people died from coronary heart disease (CHD) in 2015, which is 13% of all deaths that year across the globe.

In the late 1940s, the U.S. government took steps to study cardiovascular disease. In order to develop high quality data for their study, they decided to track a large cohort of initially-healthy people over time. The town of Framingham, Massachusetts (a suburb of Boston) was selected as the site for the study, which commenced in 1948. The study enrolled 5,209 participants aged 30-62. Participants were given a questionnaire and a medical exam every two years. They also collected data on the participants' physical characteristics and behavioral characteristics, in addition to the medical test data. Over the years, the study has expanded to include multiple generations and has collected many more factors including genetic information. This data is now famously known and is simply called the Framingham Heart Study.

In this exercise, we aim to build models using Framingham Heart Study data in order to predict CHD and to make recommendations to better prevent heart disease. The dataset is in the file framingham.csv. There are 3,658 observations, with each observation representing the data from a particular study participant. There are 16 variables in the dataset, which are described in Table 1.We want to predict TenYearCHD (whether the patient experiences coronary heart disease within 10 years of their first examination). As a consequence of your modeling efforts, we should be able to identify risk factors, which are the variables that increase the risk of CHD.

Data and EDA

Table 1: Variables in the dataset `framingham.csv`.

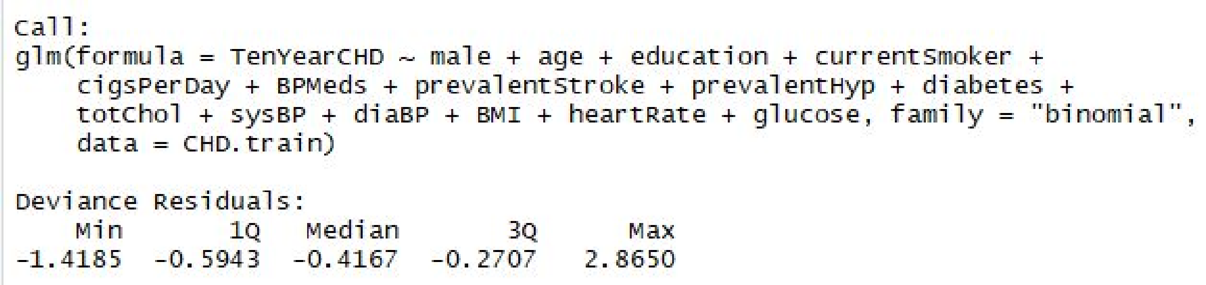


Variable	Description
`male`	Is biological sex male
`age`	Age (in years) at first examination
`education`	Some high school, high school/GED, some college/vocational school, college
`currentSmoker`	Is a current smoker
`cigsPerDay`	Number of cigarettes per day
`BPMeds`	Is on blood pressure medication at time of first examination
`prevalentStroke`	Previously had a stroke
`prevalentHyp`	Currently hypertensive
`diabetes`	Currently has diabetes
`totChol`	Total cholesterol (mg/dL)
`sysBP`	Systolic blood pressure
`diaBP`	Diastolic blood pressure
`BMI`	Body Mass Index, weight (kg)/height (m)^2
`heartRate`	Heart rate (beats/minute)
`glucose`	Blood glucose level (mg/dL)
`TenYearCHD`	Experienced coronary heart disease within 10 years of first examination

To lower the risk of CHD, physicians can prescribe preventive medication such as blood-pressure-lowering or cholesterol-lowering medications. Many policy makers, when recommending certain preventive medications to patients at risk of developing CHD, rely on evidence-based analysis that weighs the pros and cons of such interventions. Health economic evaluation is a commonly applied methodology for decision-making that takes both medical costs and health benefits (a monetized version of improved life longevity) into consideration. In fact, many countries establish clinical practice guidelines using such formalized health economic evaluation methodologies (the National Institute for Health and Clinical Excellence in England, for example).

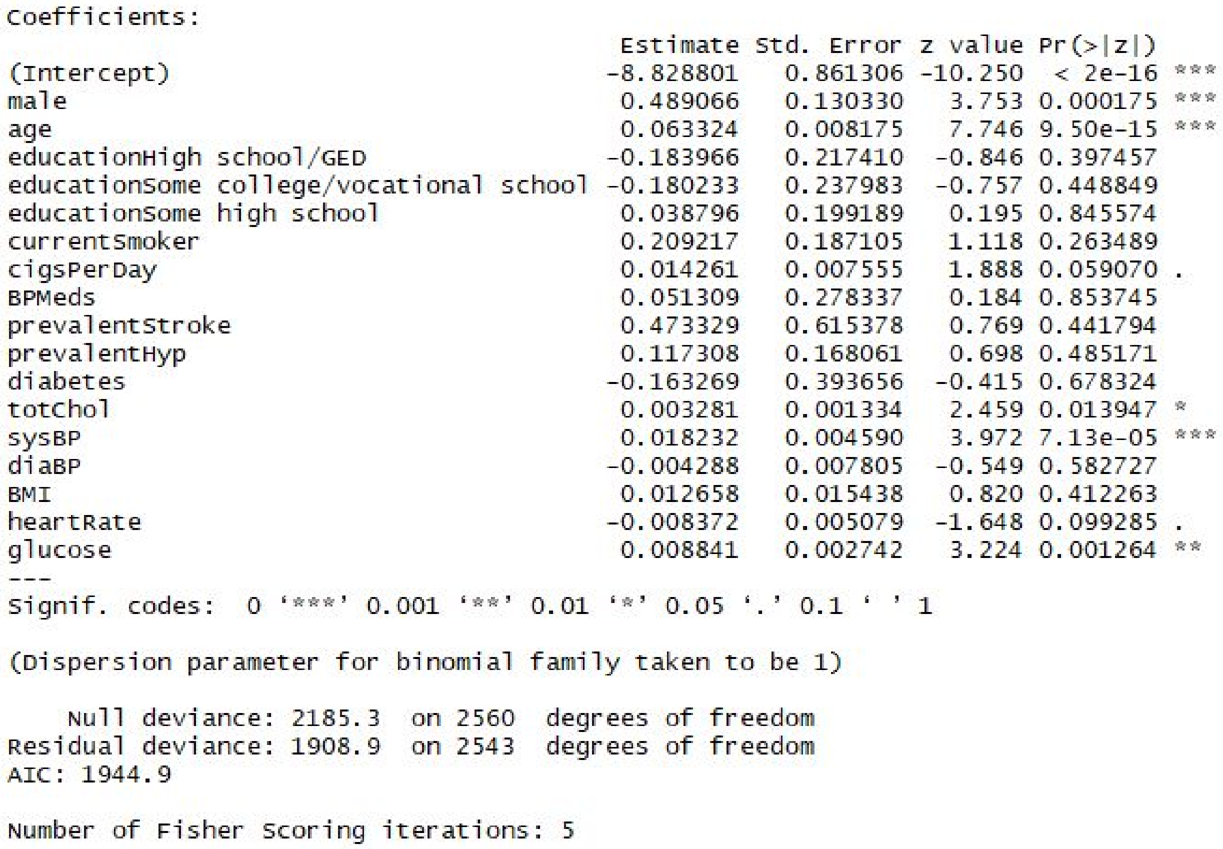
Initial Model, Logistic Regression

Using all of the provided independent variables, we build a logistic regression model to predict the probability that a patient will experience CHD within the next 10 years. Using a randomly selected subset of 70% of the data to train your model. We ensure that training and test sets have approximately equal proportions of people with TenYearCHD to people without TenYearCHD (i.e., stratified/proportional sampling).



```
Call:
glm(formula = TenYearCHD ~ male + age + education + currentSmoker +
    cigsPerDay + BPMeds + prevalentStroke + prevalentHyp + diabetes +
    totChol + sysBP + diaBP + BMI + heartRate + glucose, family = "binomial",
    data = CHD.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4185  -0.5943  -0.4167  -0.2707   2.8650
```



```
Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -8.828801   0.861306 -10.250  < 2e-16 ***
male                                    0.489066   0.130330   3.753 0.000175 ***
age                                     0.063324   0.008175   7.746 9.50e-15 ***
educationHigh school/GED               -0.183966   0.217410  -0.846 0.397457
educationSome college/vocational school -0.180233   0.237983  -0.757 0.448849
educationSome high school               0.038796   0.199189   0.195 0.845574
currentSmoker                           0.209217   0.187105   1.118 0.263489
cigsPerDay                              0.014261   0.007555   1.888 0.059070 .
BPMeds                                  0.051309   0.278337   0.184 0.853745
prevalentStroke                         0.473329   0.615378   0.769 0.441794
prevalentHyp                            0.117308   0.168061   0.698 0.485171
diabetes                               -0.163269   0.393656  -0.415 0.678324
totChol                                 0.003281   0.001334   2.459 0.013947 *
sysBP                                   0.018232   0.004590   3.972 7.13e-05 ***
diaBP                                  -0.004288   0.007805  -0.549 0.582727
BMI                                     0.012658   0.015438   0.820 0.412263
heartRate                              -0.008372   0.005079  -1.648 0.099285 .
glucose                                 0.008841   0.002742   3.224 0.001264 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2185.3  on 2560  degrees of freedom
Residual deviance: 1908.9  on 2543  degrees of freedom
AIC: 1944.9

Number of Fisher Scoring iterations: 5
```

We have 14 independent variables and 3 categorical factor variables coming from the 15th independent variable:

$$\beta TX = -8.828801 + 0.489066X1 + 0.063324X2 - 0.183966X3 + 0.180233X4 + 0.038796X5 + 0.209217X6 + 0.014261X7 + 0.051309X8 + 0.473329X9 + 0.117308X10 - 0.163269185X11 + 0.003280885X12 + 0.018231811X13 - 0.004287863X14 + 0.012658135X15 - 0.008372146X16 + 0.008841201X17$$

Where the indices 1 to 17 represent the set of following variables, in order:

1. male, 2. age, 3. education HighSchool/GED, 4. educationSome college/vocational school, 5. educationSome high school, 6. currentSmoker, 7. cigsPerDay, 8. BPMeds, 9. prevalentStroke, 10. prevalentHyp, 11. diabetes, 12. totChol, 13. sysBP, 14. diaBP, 15. BMI, 16. heartRate, 17. glucose

Feature Significance

The most important risk factors identified by the model are: Male, Age, CigsperDay, totChol, sysBP, heartRate, glucose, all significant up to 5 percent or lower.
We know that the odds of a certain event is:

$$Odds[P(Y|X)] = P(Y|X)\ 1 - P(Y|X)$$

In this case, we also know that:

$$log(Odds[P(Y|X)]) = \hat{}\beta 0 + \hat{}\beta 1H + \hat{}\beta 2X2 + ... + \hat{}\beta 17X17$$

Having said that, the variable Age is highly significant, we know that a 1 unit increase in age (so being only 1 year older) increases the odds of developing CHD by $e\hat{}\ \beta 2 = e0.063324 = 1.065$ times! With only a 10 years age difference, the odds of developing CHD doubles. We can see then that the current age of the patient has a significant, increasing impact on his future CHD condition in the next 10 years, which is completely logical.

Evaluation

We compute the results of the confusion matrix obtained by fitting the model on the test set, with a probability threshold of $p = 0.16$:

True Negatives = 677	True Positives = 111
False Negatives = 56	False Positives = 253

Thus the Accuracy, TPR and FPR can be computed:

The Accuracy describes the fraction of correct prediction ”guesses”, comparing the prediction (with the threshold used) to reality.

$$\mathit{Accuracy} = \mathit{NumberCorrect} / \mathit{NumberTotal} = 677+111/1097 = 0.718$$

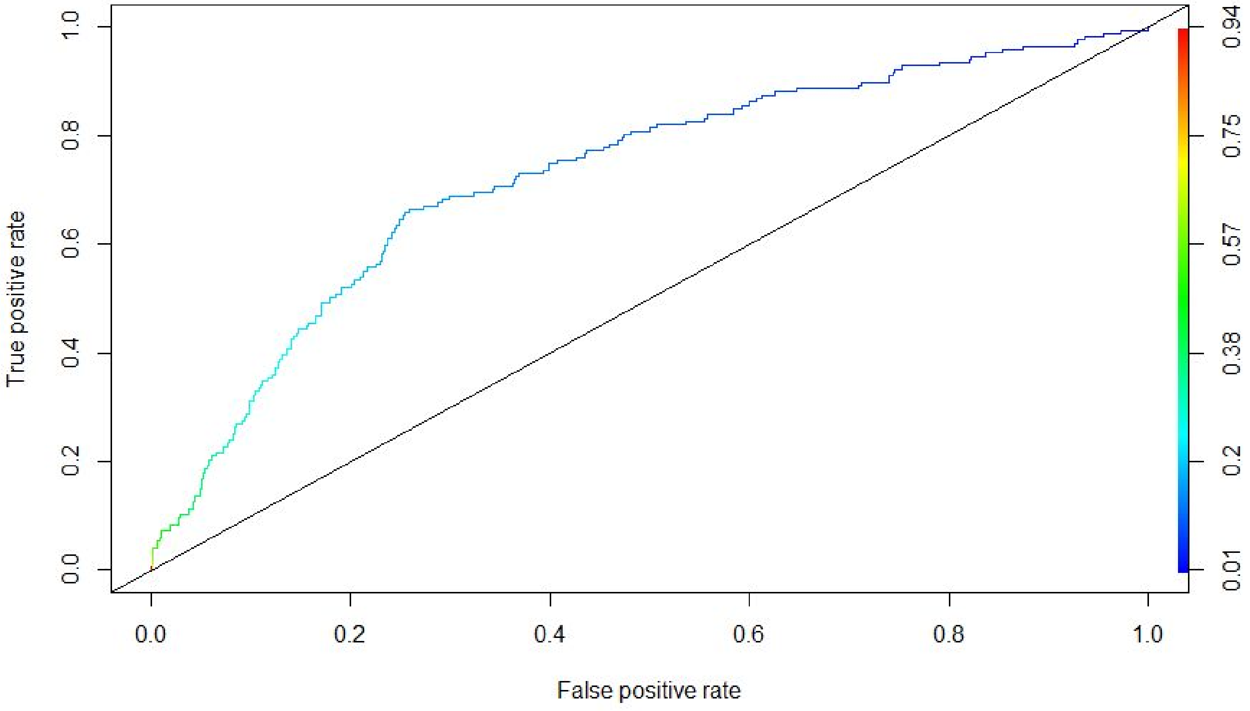
The True Positive Rate describes the fraction of correct Positive guesses (patients to which we decided to prescribe medication and that actually developed CHD) to the number of all actual people who developed CHD:

$$\mathit{TPR} = \mathit{TruePositives}/\mathit{AllPositives} = 111/111+56 = 0.665$$

The False Positive Rate describes the fraction of patients to which we wrongly prescribed medication as we thought they were high risk (but did not develop CHD), to all of those who did not develop CHD.

$$\mathit{FPR} = \mathit{FalsePositives}/\mathit{AllNegatives} = 253/253+677 = 0.272$$

ROC Curve



The ROC curve lets us visualize how the metric developed before (TPR, FPR) vary with the threshold p set. This means that we can gauge the performance of the model, regardless of a certain chosen threshold, which comes from an individual economic analysis. This gives a universal metric.

The model has an AUC of 0.725 for the Test Set. We can see that the model performs overall relatively well, and is better than the baseline. The AUC gives us the likelihood that the model would assign a higher CHD probability to the patients who will develop CHD.

We can see from the curve that there is definitely some improvements to be made, especially with the True Positive Rate, which is lagging at the beginning of the curve. This is representative of the results obtained beforehand, where we could actually see that the amount of True Positives, which we most care about, was low. This results is also translated in a low accuracy for the model.

To compare this model with other models across multiple thresholds (removing the threshold dependency), the ROC curve is very useful in that it provides a simple and visual metric for directly comparing. We could build many other predictive models using other medications and see if they actually improve on the costs, accuracy (and amount of True Positives) and ultimately plot the ROC curve, and gauge the results.

CODE

```
library(dplyr)
library(ggplot2)
library(GGally)
library(caTools)
library(ROCR)
library(MASS)

# Read data- Framingham file
CHD <- read.csv("framingham.csv")
str(CHD)
head(CHD)

# No need to worry about factors for Logistic Regression
# Split into train and test
set.seed(142)
split = sample.split(CHD$TenYearCHD, SplitRatio = 0.7)
CHD.train <- filter(CHD, split==TRUE)
CHD.test <- filter(CHD, split==FALSE)

# How many people have had CHD ?
table(CHD.train$TenYearCHD)
table(CHD.test$TenYearCHD)

# Baseline Model: predict that no one will have CHD in 10 years

#training baseline accuracy
baseline_accuracy_tr = 2171/(2171+390)
#test baseline accuracy
baseline_accuracy_ts = 930/(930+167)

# Fit logistic regression model
model <- glm(TenYearCHD ~ male + age + education + currentSmoker +
             cigsPerDay + BPMeds + prevalentStroke + prevalentHyp +
             diabetes + totChol + sysBP + diaBP + BMI + heartRate +
             glucose, data=CHD.train,family = "binomial")

summary(model)

#test set predictions
```

```
test.pred = predict(model, newdata=CHD.test, type="response")
summary(test.pred)

#confusion matrix, threshold = 0.16
table(CHD.test$TenYearCHD, test.pred > 0.16)

# predict a single value for following clinic patient:
# Female, age 51, college education, currently a smoker with an average of
# 20 cigarettes per day. Not on blood pressure medication, has not had stroke,
# but has hypertension. Not diagnosed with diabetes; total Cholesterol at
# 220. Systolic/diastolic blood pressure at 140/100, BMI at 31, heart rate
# at 59, glucose level at 78.


CHD.obs <- data.frame(male=0, age=51, education = 'College', currentSmoker = 1,
                cigsPerDay = 20, BPMeds = 0, prevalentStroke = 0, prevalentHyp =1,
                diabetes =0, totChol=220, sysBP=140, diaBP=100, BMI=31,
                heartRate=59, glucose=78)
predict(model, newdata=CHD.obs, type="response")

# ROC curve
rocr.log.pred <- prediction(test.pred , CHD.test$TenYearCHD)
logPerformance <- performance(rocr.log.pred, "tpr", "fpr")
plot(logPerformance, colorize = TRUE)
abline(0, 1)

as.numeric(performance(rocr.log.pred, "auc")@y.values)
```