Nicolas Kardous

IND 242 Homework 2

Problem 3

**Part a)**

# IND242HW2 Problem 3

# Nicolas Kardous

library(dplyr)

library(ggplot2)

library(GGally)

library(caTools)

library(ROCR)

# Part a)

mydata = read.csv("framingham.csv")

# i)

split = sample.split(mydata$TenYearCHD, SplitRatio = 0.7)

mydata.train <- filter(mydata, split == TRUE)

mydata.test <- filter(mydata, split == FALSE)

mod <- glm(TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay + BPMeds

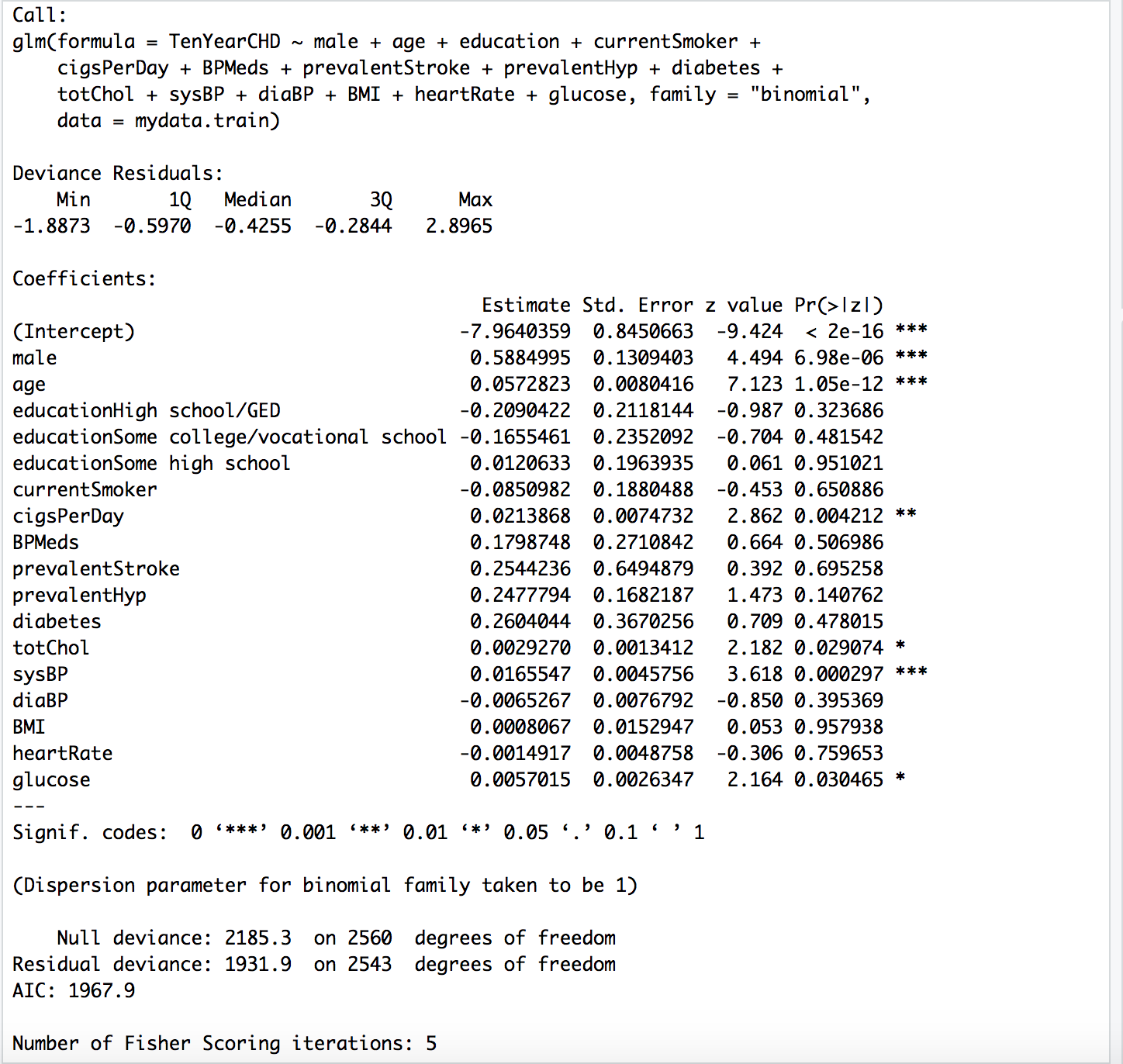
+ prevalentStroke + prevalentHyp + diabetes + totChol + sysBP + diaBP

+ BMI + heartRate + glucose, data = mydata.train, family = "binomial")

# ii)

summary(mod)

Looking at the values for beta, we find significant values of beta for male, age, educationHigh school/GED, totChol, sysBP and glucose. I believe totChol is very correlated to TenYearCHD because cholesterol affects a persons blood pressure, which will also affect a persons risk of heart disease. When there is more cholestorol, it builds up the walls in the arteries, increasing the risk of heart disease. Because the beta coefficient for totChol is 0.003109, then the odds is e^0.003109 = 1.003



# iii)

Threshold value of p is 0.16 such that it is optimal to perscribe the medication to a patient if. their 10 year CHD exceeds p

# iv)

predTest = predict(mod, newdata=mydata.test, type="response")

summary(predTest)

table(mydata.test$TenYearCHD, predTest > 0.16)

Model Accuracy: (648+108)/(648+282+59+108) = 0.69. Given this, our model is some what accurate because this value shows that our model is accurate 69% of the time. The accuracy was found to be the proportion of the TenYearCHD that was correctly classified vs the TenYearCHD that was incorrectly classified

True Positive Rate (TPR): 108/(59+108) = 0.6467. The TPR is our estimate of the conditional probability that that our classifier makes a correct prediction given Y = 1. Proportion of patients that have CHD that we correctly prescribed with the medication

False Positive Rate (FPR): 282/(648+282) = 0.303. The FPR is the proportion of patients that do not have CHD that we incorrectly prescribed the medicine for

> predTest = predict(mod, newdata=mydata.test, type="response")

> summary(predTest)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.01560 0.06473 0.12117 0.15220 0.19861 0.73069

> table(mydata.test$TenYearCHD, predTest > 0.16)

FALSE TRUE

0 656 274

1. 60 107

# v)

Cost\_per\_patient <- ((560000\*(0.16/4))+(60000\*(1-(0.16/4))))/(282+108)

Cost per patient is around $205.128

This assumption is not reasonable because this assumption believes that the treatment decision is independent of a patient's risk of developing CHD. However, there is the possiility that the treatment decision could impact a patients risk of developing CHD

Cost\_per\_patient\_adj <- ((560000\*(0.16/4))+(60000\*(1-(0.16/4)))+(500000\*0.16)+(0\*(1-0.16)))/(648+282+59+108)

Adjusted cost per patient is around $145.852

# vi)

# Baseline model: predict that no one defaults

# Accuracy of baseline on testing:

table(mydata.test$TenYearCHD)

Model Accuracy: (930+0)/(930+0+167+0) = 0.848. Given this, our baseline model is accurate because this value shows that our baseline model is accurate 84.8% of the time. The accuracy was found to be the proportion of people who actually don't need treatment vs the total people who don't and do need treatment

True Positive Rate (TPR): 0/(0+167) = 0. The TPR is our estimate of the conditional probability that the patients who do require the medication vs the patients who do and don't need the treatment

False Positive Rate (FPR): 0/(0+930) = 0. The FPR is the proportion of patients that actually do not have CHD vs the total number of patients who are perscribed and not perscribed medication, but actually don't have CHD

Cost\_per\_patient\_vi <- (500000\*0.16)+(0\*(1-0.16))/(930+167)

Cost per patient here is $80,000

# vii)

vii <- 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)

predTest\_vii = predict(mod, newdata=vii, type="response")

If we have the following information about the 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.

The predicted probability that the patient will experience CHD in the next year is 0.147 or 14.7%. Because our threshold from the decision tree is 0.16, the physician should not prescribe the preventative medication to the patient

# Part b)

rocr.log.pred <- prediction(predTest, mydata.test$TenYearCHD)

logPerformance <- performance(rocr.log.pred, "tpr", "fpr")

plot(logPerformance, colorize = TRUE)

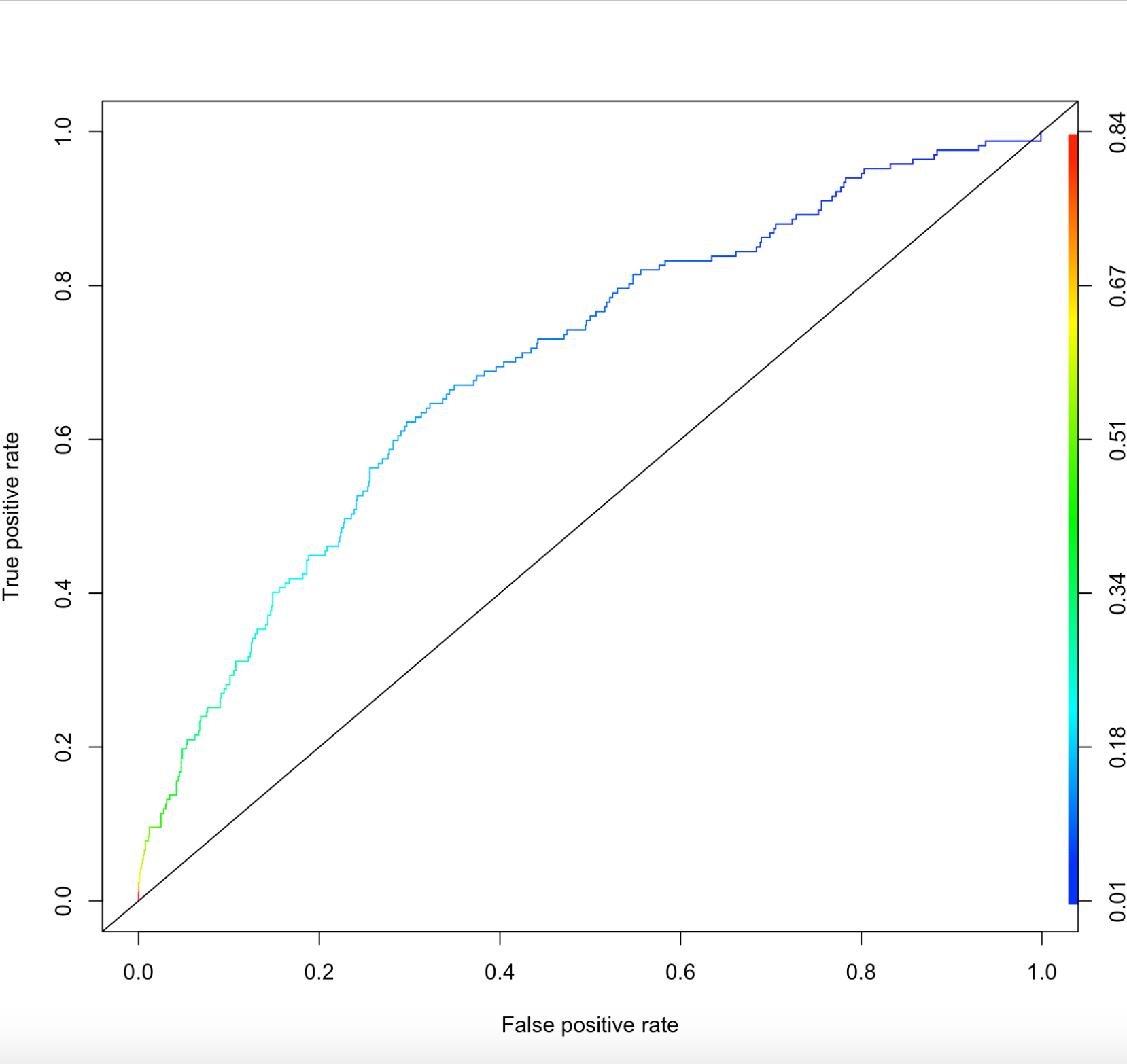
abline(0, 1)

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

The ROC curve plots the TPR and FPR for every break-even threshold p between 0.0 and 1.0 n ROC curve can also be drawn for LDA, and for any method that predicts probabilities

This curve is helpful to decision makers looking to further study the medication for CHD as well as other medications because these decision makers could find a suitable breakeven threshold. One interesting observation implied by the ROC curves is that we want to simultaneously achieve a high TPR and low FPR n. This corresponds to a high area under the ROC curve, which is called the AUC (Area Under the Curve)

The AUC value we found for our ROC curve is 0.7588629



# Part c)

The insurance company should charge a price of $47,368.4211

# Part d)

Some ethical concerns raised by the analysis is that it is assumed that there is no negative side effect for people who don't have CHD but who take the medicine. In reality, there could be a negative side effect. Additionally, a threshold value of 16% would seem pretty low for someone to be deemed to require medication. Thus, one way this analysis could be changed in the future, is to increase the threshold value from 16%, to increase the chance for people who have CHD to receive medication.