Homework 2

Problem 1:

a) Probability that a student who studies for 40 has an undergrad GPA of 3.5 gets an A in the class.

P(Y=1/X=x) = 1/(1 + exp(-(-6 + 0.05*40 + 1*3.5))) = 0.38

b) Find the number of hours

$$p = \frac{1}{1+e^{-A}} \rightarrow A = \log\left(\frac{p}{1-p}\right) \ with\ A = \ -6 + 0.05 * X1 + 1 * 3.5 = 0 \ \rightarrow X1 = 50\ hours$$

Problem 2:

$$P(\,D \mid X = 4\,) = \frac{f_{X|D} * P(D)}{f_X} = \frac{f_{X|D} * P(D)}{f_{X|D} * P(D) + f_{X|D} * P(D^C)} \text{ with } \begin{cases} \mathrm{P(D)} = \ 0.8, \\ \mathrm{P(D^C)} = 0.2 \\ f_{X/D} = \frac{1}{\sqrt{2\pi * 36}} e^{-\frac{(mean-10)^2}{2*36^2}} \end{cases}$$

$$P(\,D \mid X = 4\,) = 0.75$$

Problem 3: Framingham Heart Study

a)

i) Fitted logistic model

P(Y=1/X=x) = 1/(1 + exp(-(0.43*male + 0.064*age -0.121*educationHighSchool -0.077* educationSomeCollege + 0.064*educationSome High school + 0.103*currentSmoker + 0.017*cigsPerDay - 0.107*BPMeds + 0.936*prevalentStroke + 0.244*prevalentHyp - 0.0059*diabetes + 0.0018*totChol + 0.0167*sysBP – 0.0074*diaBP + 0.004*BMI - 0.0000008*heartRate + 0.0083*glucose -8.49)))

ii) The most important risk factors

The most important (significant with p.value<0.05) risk factors: **male, age, cigsPerDay, sysBP, glucose**.

Interpretation: Holding all other variables constant, being a man yields a $e^{0.43}$ times higher estimated odds of having a CHD

iii) Identify a threshold

(p/4)560000+60000-15000p = 500000p

p = 0.16

iv) Test set performance on the logistic model

	Good risk P(Y=1)<0.16	**Bad risk P(Y=1)>0.16**
No CHD	**637**	**293**
CHD	**56**	**111**

$$Accuracy = \frac{637 + 111}{1097} = 0.681$$

Accuracy: The model will effectively predict people that will or will not have CHD for 68.2% of the patients

$$Tue\ Positive\ Rate = \frac{111}{56 + 111} = 0.664$$

True Positive Rate: Among the people who will actually have a CHD, the algorithm accuretaly predict 66.4% of that population

$$False\ Positive\ Rate = \frac{293}{293 + 637} = 0.315$$

False positive Rate: Among the people who will actually not have a CHD, the algorithm was wrong about 31.5% of that population.

v) Expected costs

	Good risk P(Y=1)<0.16	**Bad risk P(Y=1)>0.16**
No CHD	**0**	**60000**
CHD	**500000**	**560000**

Expected cost per patient: $Cost = \frac{(293)*60000+560000*111+56*500000}{1097} = 98213$

It is a fair assumption to assume that the CHD outcomes in our test set are not affected by the treatment decisions because ou data was collected before the medication was considered. Which means it cannot have been affected their lifestyle thus the probability of their positiveness.

If the treatment decision influence the CHD outcome thus we would have actually 75% less people positive

	Good risk P(Y=1)<0.16	**Bad risk P(Y=1)>0.16**
No CHD	**637**	**293+84**
CHD	**56**	**111-84**

New expected cost per patient: $Cost = \frac{(376)*60000+560000*(28)+500000*56}{1097} = 59927$

vi) Baseline model that predicts none of the patients are high risk for CHD

	Good risk P(Y=1)<0.16	**Bad risk P(Y=1)>0.16**
No CHD	**930**	**0**
CHD	**167**	**0**

It will be correct on the 930 people that are negative

$$Accuracy_{Baseline\ Model} = \frac{930}{1097} = 0.85$$

$$Tue\ Positive\ Rate = 0$$

$$False\ Positive\ Rate = 0$$

Baseline expected cost per patient: $Cost = \frac{930*0+167*500000}{1097} = 76117$

Better accuracy, more costs

vii) New patient

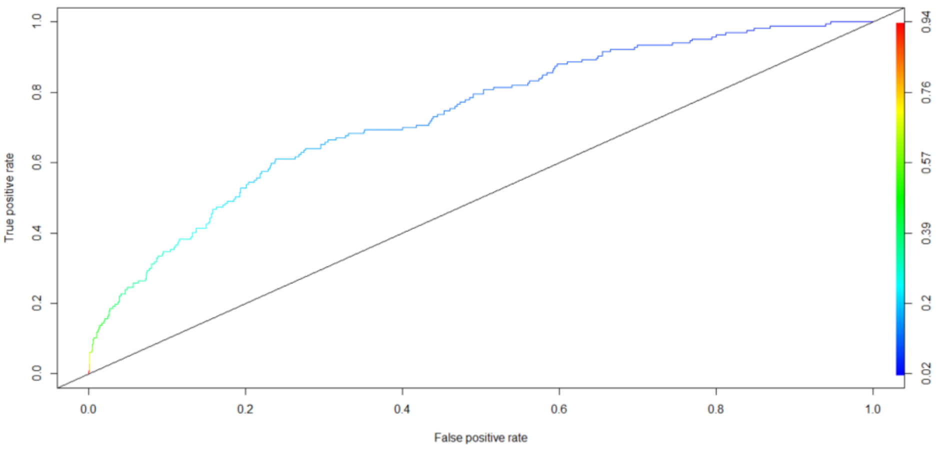
Data: Female, age 51, college education, currently a smoker with an average of 20 cigarettes per day. Not on blood pressure medication, has not had stroke, but has hypertension. Not diagnosed with diabetes; total Cholesterol at 220. Systolic/diastolic blood pressure at 140/100, BMI at 31, heart rate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

P(Y=1/X=x) = 1/(1 + exp(-(0.5*0 + 0.066*51 -0.273*0 -0.160* 1 -0.037*0 + 0.186*1 + 0.017*20 + 0.152*0 + 0.522*0 + 0.247*1 + 0.045*0 + 0.002*220 + 0.014*140 – 0.003*100 + 0.005*31 - 0.0003*59 + 0.009*78)))

P(Y=1/X=x) = 0.157

P < 0.16 the physician should not prescribe the medication for this patient.

b) ROC Curve



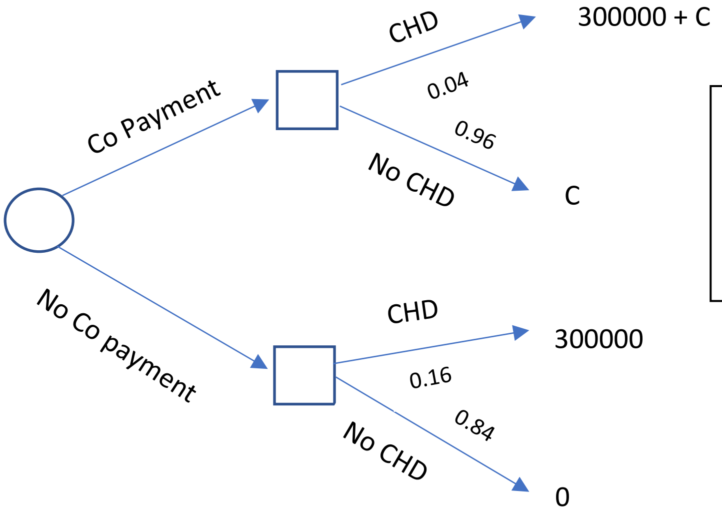
The ROC curve allow the physicians to wisely choose the break-even threshold regarding the FPR and TPR they intend to achieve. It can also be useful to compare this curve with those of other medications to select those with the higher AUC

Observation: the TPR increase with the FPR when p decrease. Something interesting about the graph is that, right before FPR = 0.4, FPR increase while the TPR is almost constant and with small values of values of p, the ratio TPR/FPR globally diminish (right side of the curve)

The AUC is the likelihood that the model would assign a higher CHD probability to the customer who actually will be positive

AUC = 0.7335

c) Health insurance



We assume that all the patients are insured, and we are only thinking whether they will be willing to pay the extra C for the disease's prevention

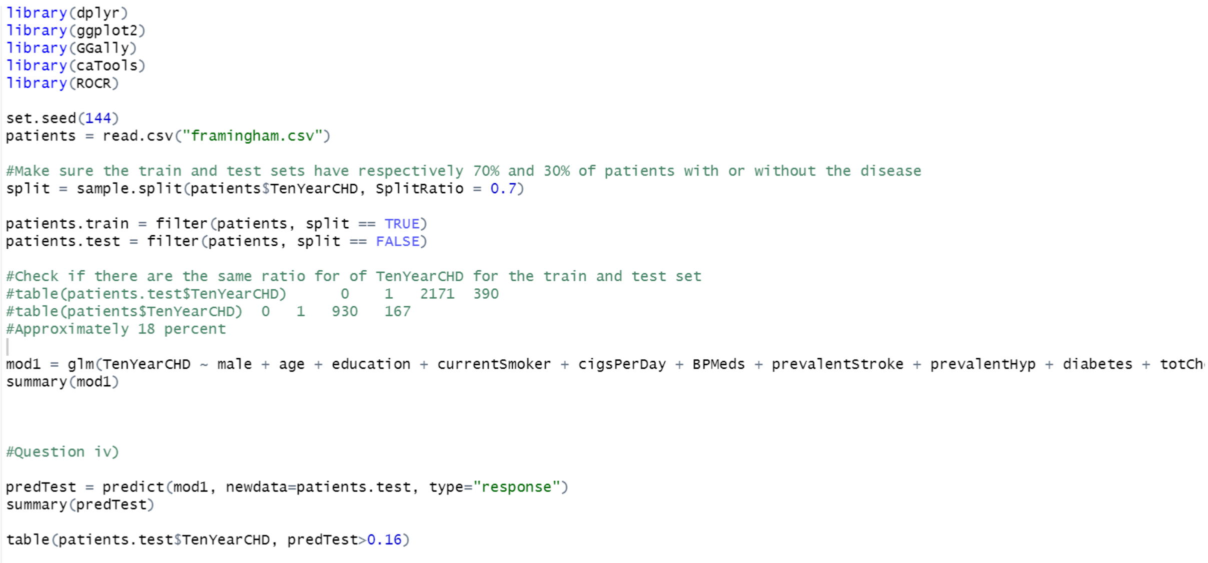
$$0.04 * (300000 + C) + 0.96 * C = 300000 * 0.96$$

$$C = 36000$$

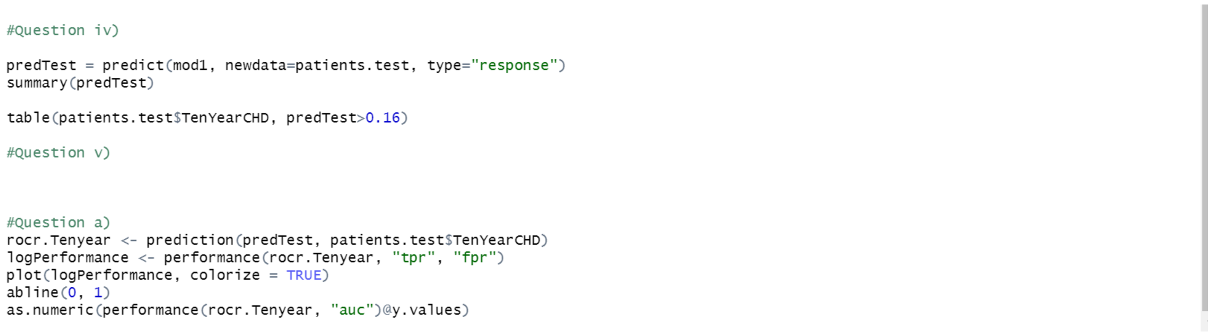
d) Model improvement

The fact that we decide whether or not a patient will be positive to an illness according to the expected cost for an insurance company raise some ethical issues. According to the ROC curve, we could have a better TPR even though it would increase the FPR. The consequence would be good for the patients, but less profitable for the Insurance. They can tend to lower their costs.

CODE



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

set.seed(144)
patients = read.csv("framingham.csv")

#Make sure the train and test sets have respectively 70% and 30% of patients with or without the disease
split = sample.split(patients$TenYearCHD, SplitRatio = 0.7)

patients.train = filter(patients, split == TRUE)
patients.test = filter(patients, split == FALSE)

#Check if there are the same ratio for of TenYearCHD for the train and test set
#table(patients.test$TenYearCHD)      0    1   2171  390
#table(patients$TenYearCHD)  0   1   930   167
#Approximately 18 percent

mod1 = glm(TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay + BPMeds + prevalentStroke + prevalentHyp + diabetes + totCh
summary(mod1)



#Question iv)

predTest = predict(mod1, newdata=patients.test, type="response")
summary(predTest)

table(patients.test$TenYearCHD, predTest>0.16)
```



```
#Question iv)

predTest = predict(mod1, newdata=patients.test, type="response")
summary(predTest)

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

#Question v)


#Question a)
rocr.Tenyear <- prediction(predTest, patients.test$TenYearCHD)
logPerformance <- performance(rocr.Tenyear, "tpr", "fpr")
plot(logPerformance, colorize = TRUE)
abline(0, 1)
as.numeric(performance(rocr.Tenyear, "auc")@y.values)
```