

author: 3032247297

2)

```
library(readr)
library(caTools)
df <- read_csv("C:/Users/Murtz.Kizilbash/Desktop/ieor142/hw2/framingham.csv")
```

```
## Parsed with column specification:
## cols(
##   male = col_double(),
##   age = col_double(),
##   education = col_character(),
##   currentSmoker = col_double(),
##   cigsPerDay = col_double(),
##   BPMeds = col_double(),
##   prevalentStroke = col_double(),
##   prevalentHyp = col_double(),
##   diabetes = col_double(),
##   totChol = col_double(),
##   sysBP = col_double(),
##   diaBP = col_double(),
##   BMI = col_double(),
##   heartRate = col_double(),
##   glucose = col_double(),
##   TenYearCHD = col_double()
## )
```

```
set.seed(123)
```

```
sample = sample.split(df$male, SplitRatio = .7)
train = subset(df, sample == TRUE)
test = subset(df, sample == FALSE)
```

i)

```
model <- glm(TenYearCHD ~ glucose + heartRate + BMI + diaBP + sysBP + totChol + diabetes + prevalentHyp +
summary(model)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ glucose + heartRate + BMI + diaBP +
##   sysBP + totChol + diabetes + prevalentHyp + prevalentStroke +
##   BPMeds + cigsPerDay + currentSmoker + age + education + male,
##   data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
```

```

## -0.70155 -0.18441 -0.10427 -0.01103 1.10208
##
## Coefficients:
##
## Estimate Std. Error t value
## (Intercept) -6.399e-01 9.348e-02 -6.846
## glucose 9.817e-04 3.838e-04 2.558
## heartRate -3.124e-04 5.818e-04 -0.537
## BMI 1.911e-03 1.832e-03 1.043
## diaBP -2.158e-03 9.806e-04 -2.201
## sysBP 3.462e-03 5.901e-04 5.867
## totChol 7.501e-05 1.604e-04 0.468
## diabetes 2.686e-02 5.204e-02 0.516
## prevalentHyp 6.274e-03 2.077e-02 0.302
## prevalentStroke 7.060e-02 9.130e-02 0.773
## BPMeds 1.889e-02 4.241e-02 0.445
## cigsPerDay 2.126e-03 9.208e-04 2.309
## currentSmoker 2.200e-02 2.164e-02 1.017
## age 6.581e-03 9.233e-04 7.127
## educationHigh school/GED -9.063e-03 2.374e-02 -0.382
## educationSome college/vocational school -4.464e-03 2.635e-02 -0.169
## educationSome high school 1.376e-02 2.311e-02 0.595
## male 5.491e-02 1.485e-02 3.697
## Pr(>|t|)
## (Intercept) 9.49e-12 ***
## glucose 0.010591 *
## heartRate 0.591336
## BMI 0.297195
## diaBP 0.027821 *
## sysBP 5.02e-09 ***
## totChol 0.640008
## diabetes 0.605791
## prevalentHyp 0.762611
## prevalentStroke 0.439435
## BPMeds 0.656060
## cigsPerDay 0.021030 *
## currentSmoker 0.309456
## age 1.33e-12 ***
## educationHigh school/GED 0.702712
## educationSome college/vocational school 0.865498
## educationSome high school 0.551607
## male 0.000223 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1146586)
##
## Null deviance: 325.00 on 2559 degrees of freedom
## Residual deviance: 291.46 on 2542 degrees of freedom
## AIC: 1740.5
##
## Number of Fisher Scoring iterations: 2

```

$tenyearCHD = -.06 - .0003(heartRate) - .002(diaBP) + .00008(totChol) + .0063(prevalentHyp) + .01(BPmeds) + .022(curren$

- ii) According to the summary of the significance of the features, the most important risk factors in predicting whether or not someone will have CHD in 10 years is their age. When it comes to age, every increase in age increases the log odds of 10yearCHD by .007

iii)

$$560000(p/4) + 60000(1 - p/4) = 500000(p)$$

$$p = .16$$

iv)

```
test$prediction = predict(model, newdata = test, type = 'response')

high_risk <- subset(test, prediction >= .16)
low_risk <- subset(test, prediction < .16)

tp = nrow(subset(test, prediction >= .16 & TenYearCHD == 1))
fn = nrow(subset(test, prediction < .16 & TenYearCHD == 1))
fp = nrow(subset(test, prediction >= .16 & TenYearCHD == 0))
tn = nrow(subset(test, prediction < .16 & TenYearCHD == 0))

tpr = tp/(tp + fn)
fpr = fp/(fp+tn)
accuracy = (tp+tn)/(tp+fp+fn+tn)

tpr
```

```
## [1] 0.68
```

```
fpr
```

```
## [1] 0.3770314
```

```
accuracy
```

```
## [1] 0.6320583
```

the true positive rate is .68 the false positive rate is .377 the accuracy is .63

The tpr tells us the number of people who contracted CHD in 10 years that were correctly identified.

The fpr tells us the proportion of negative cases incorrectly identified by the model

the accuracy tells us the proportion of the data that was correctly identified.

v)

if chd is not affected by treatment:

$$EXPECTEDCOST = \frac{36(500000) + 131(560000) + 423(60000)}{1507 + 423 + 131 + 136}$$

Which equals 106417.5 dollars.

This assumption does not make much sense because if taking medicine does not have an affect on the development of the condition, then the premise of this study is invalid.

if taking preventative medicines does reduce the outcome of CHD:

$$EC = \frac{35(500000) + 13(.08)(560000) + 23(1.2)(60000)}{1097}$$

so the expected cost in this case is 97670

vi)

```
predTest = predict(model, test, type = 'response')
table(test$TenYearCHD, predTest > .999)
```

```
##
##      FALSE
##    0    923
##    1    175
```

vii)

```
new <- data.frame(male=0, age=51, education = 'College', currentSmoker=1, cigsPerDay = 20, BPMeds = 0, p
predict(model, newdata = new, type = 'response')
```

```
##          1
## 0.1692132
```

the predicted probability that this patient will contract CHD in the next 10 years is .17, we should prescribe the medicine bc the patient probability exceeds the threshold.

b)

```
library(plotROC)
```

```
## Loading required package: ggplot2
```

```
ggplot(test, aes(d = test$TenYearCHD, m = test$prediction)) + geom_roc()
```

