Emily Barnes
9/27/2019

Assignment #2

Chapter 4 Question 6

→ Data on students in a statistics class
$[X_1]$ = hours studied $[X_2]$ = undergrad gpa Y = receive an A
↳ logistic regression
$\beta_0 = -6$, $\beta_1 = .05$ $\beta_2 = 1$

a) Probability that student studies 40 hr. has GPA 3.5 gets an A
↳ logistic regression - equation

$$P_r(Y=1|X=x) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}}$$

→ with $X_1 = 40$ and $X_2 = 3.5$, probability of student getting an A is .378

b) Same student, how many hours of study to have 50% chance?
↳ would need to study for **50 hours**

$$.5 = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}}$$

$$\left[.5e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)} = .5\right] \ln$$

$$-(-6+.05X_1+3.5) = \ln(1) = 0$$

$$-2.5+.05X_1 = 0$$

$$.05X_1 = 2.5$$

$$\boxed{X_1 = 50}$$

Chapter 4 Problem 7

→ Predict Yes/No on dividend based on last year's profit
↳ for no, $\bar{X} = 0$ for yes, $\bar{X} = 10$
σ^2 for both = 36
→ assume X follows normal dist, predict prob that a company will give dividend given % profit has $X = 4$

→ density function for normal random variable is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$

Bayes Theorem:

$$Pr(y=1 \mid X=4) = \frac{\pi_K f_K(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

π = marginal probabilities μ = in class means
σ^2 = common variance

note: 80% companies issued dividend,

$\therefore \pi_0 = .2 \quad \pi_1 = .8$

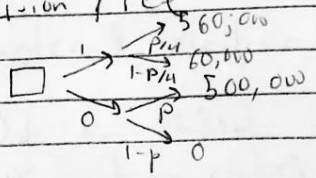
$$Pr(y=1 \mid X=4) = \frac{\pi_1 f_1(4)}{\pi_1 f_1(4) + \pi_0 f_0(4)}$$

$f_1(4) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\mu)^2/2\sigma^2}$ where $\sigma^2=36, \mu=10, X=4$
$f_1(4) = .0403$

$f_0(4) =$ " where $\sigma^2=36, \mu=0, X=4$
$f_0(4) = .0532$

$$Pr(y=1 \mid X=4) = \frac{\pi_1 f_1(4)}{\pi_1 f_1(4) + \pi_0 f_0(4)} = \frac{.8(.04)}{.8(.04)+.2(.05)} = \boxed{.7518}$$

Decision Tree



To find break even point, find cost of no prescribe = cost of prescribe

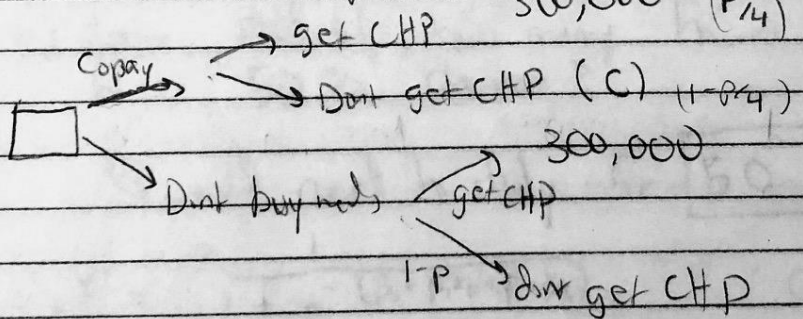
$$\therefore \quad p(500,000) = (1 - P/4)60,000 + \frac{P}{4}(560,000)$$

$$500,000p = 60,000 - 15,000p + 140,000p$$

$$500,000p = 60,000 + 125,000p$$

$$375,000p = 60,000 \qquad p = .16$$

Decision tree for patients



set the sides equal and solve for C

$$(C + 300,000).04 + C(.96) = 300,000(.16) + \cancel{0(.84)}$$

$$C + 300,000(.04) = 300,000(.16)$$

$$C = 300,000(.16 - .04) = 300,000(.12) = 36,000$$

Assignment_2_rmd

Emily

September 23, 2019

Problem 1

```
B0 = -6
B1 = .05
B2 = 1

X1 = 68
X2 = 3.5

e = exp(1)

logisticeqn = 1/(1+e^-(B0 + (B1*X1)+(B2*X2)))

logisticeqn
```

```
## [1] 0.7109495
```

Problem 2

```
pi1 = .8
pi0 = .2
pi=3.1415

sig2 = 36

u0 = 0
u1 = 10

x = 4

f1 = (1/(sqrt(2*pi*sig2)))*exp(-(((x-u1)^2)/(2*sig2)))
f1
```

```
## [1] 0.04032905
```

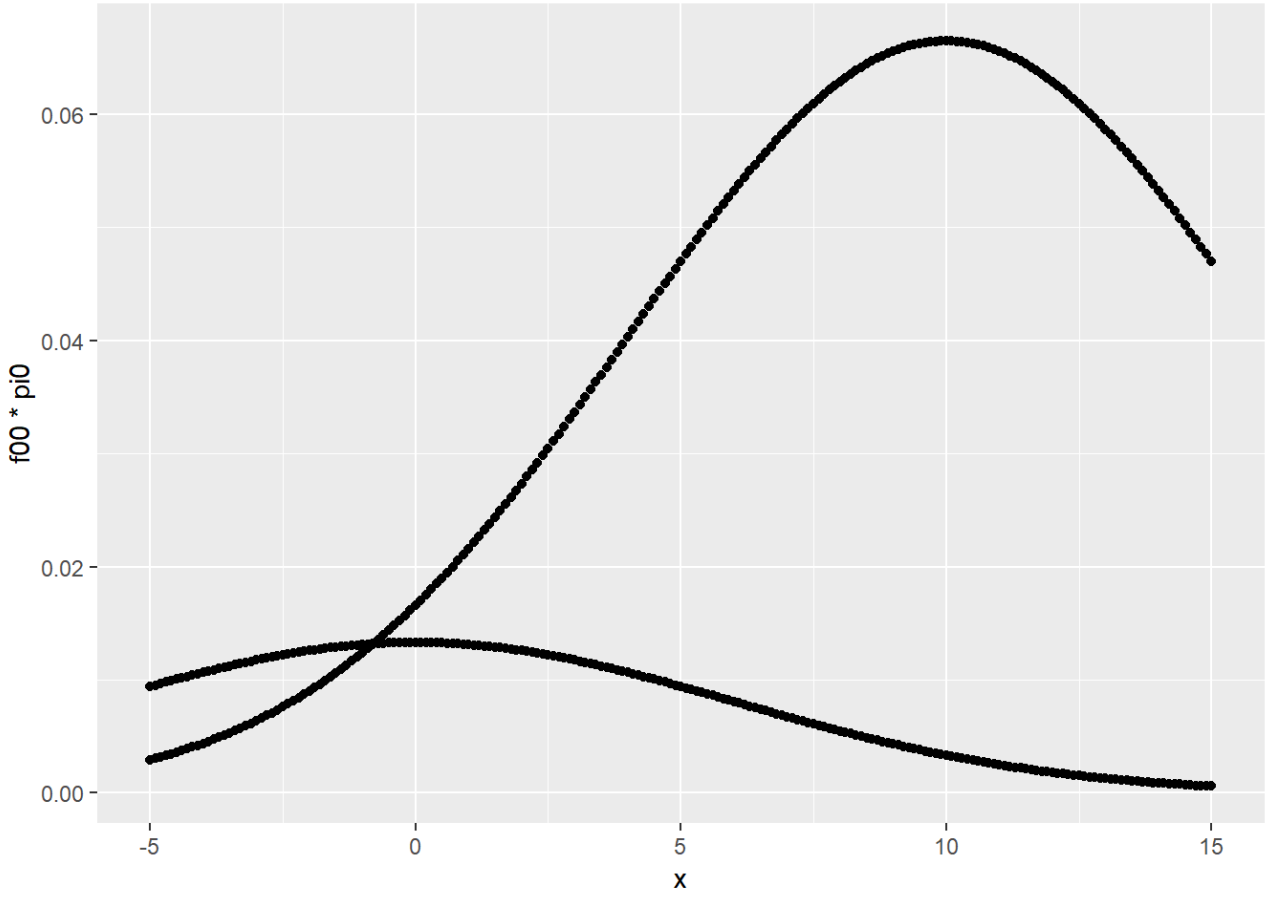
```
f0 = (1/(sqrt(2*pi*sig2)))*exp(-(((x-u0)^2)/(2*sig2)))
f0
```

```
## [1] 0.05324212
```

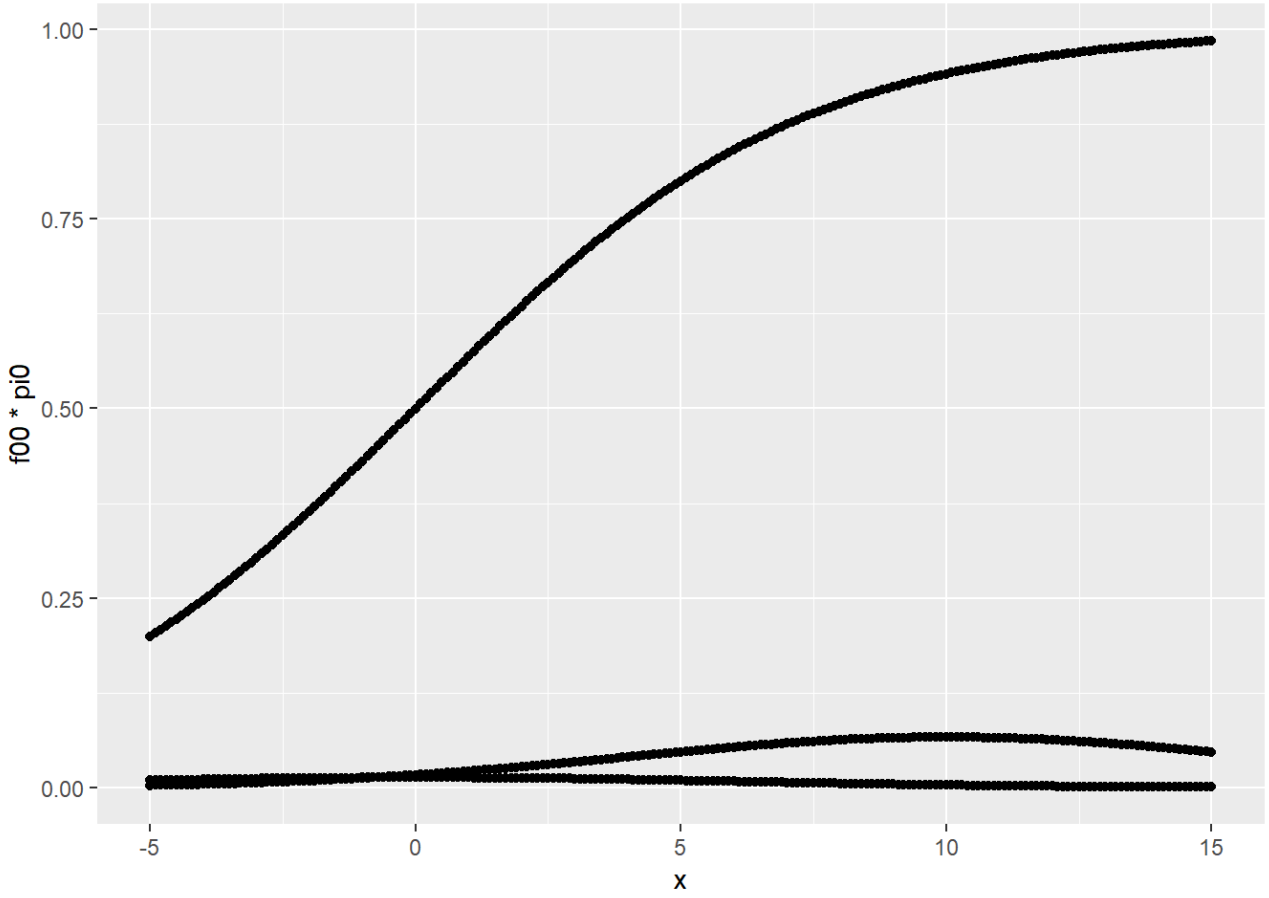
```
P0 = (pi1*f1)/((pi1*f1)+(pi0*f0))
P0
```

```
## [1] 0.7518525
```

```
x = seq(from= -5, to = 15, by = .1)

f00 = (1/(sqrt(2*pi*sig2)))*exp(-(((x-u0)^2)/(2*sig2)))
f11 = (1/(sqrt(2*pi*sig2)))*exp(-(((x-u1)^2)/(2*sig2)))

xx <- data.frame(x, f00, f11)

ggplot(data = xx, aes(x = x, y = f00*pi0))+
  geom_point()+
  geom_point(aes(x = x, y = f11), data = xx)
```



```
xx <- xx %>%
  mutate(prob1 = (pi1*f11)/((pi1*f11)+(pi0*f00)))

ggplot(data = xx, aes(x = x, y = f00*pi0))+
  geom_point()+
  geom_point(aes(x = x, y = f11), data = xx)+
  geom_point(aes(x=x, y = prob1), data = xx)
```



Problem 3

```
dat <- read_csv("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(144)

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

# what is a split?
chd.train <- filter(dat, split == TRUE) # is split a variable in loans?
chd.test <- filter(dat, split == FALSE)

table(chd.train$TenYearCHD)
```

```
##
##    0    1
## 2171  390
```

```
table(chd.test$TenYearCHD)
```

```
##
##   0   1
## 930 167
```

```
#ggscatmat(chd.train)
```

Part 3 a i

You can also embed plots, for example:

```
mod1 <- glm(TenYearCHD~., data=chd.train, family="binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = "binomial", data = chd.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5660  -0.5879  -0.4221  -0.2890   2.8298
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                             -8.495e+00  8.528e-01  -9.962
## male                                     4.292e-01  1.306e-01   3.285
## age                                      6.495e-02  7.964e-03   8.156
## educationHigh school/GED                -1.209e-01  2.131e-01  -0.567
## educationSome college/vocational school -7.719e-02  2.311e-01  -0.334
## educationSome high school                6.404e-02  1.970e-01   0.325
## currentSmoker                            1.033e-01  1.868e-01   0.553
## cigsPerDay                               1.739e-02  7.411e-03   2.346
## BPMeds                                  -1.072e-01  2.835e-01  -0.378
## prevalentStroke                          9.369e-01  5.912e-01   1.585
## prevalentHyp                             2.443e-01  1.691e-01   1.445
## diabetes                                -5.921e-03  3.903e-01  -0.015
## totChol                                  1.872e-03  1.351e-03   1.386
## sysBP                                    1.678e-02  4.791e-03   3.502
## diaBP                                   -7.463e-03  7.847e-03  -0.951
## BMI                                      4.455e-03  1.546e-02   0.288
## heartRate                               -8.383e-07  4.995e-03   0.000
## glucose                                  8.356e-03  2.721e-03   3.071
##                                         Pr(>|z|)
## (Intercept)                              < 2e-16 ***
## male                                    0.001020 **
## age                                     3.47e-16 ***
## educationHigh school/GED                0.570495
## educationSome college/vocational school 0.738329
## educationSome high school               0.745147
## currentSmoker                           0.580275
## cigsPerDay                              0.018961 *
## BPMeds                                  0.705326
## prevalentStroke                         0.113042
## prevalentHyp                            0.148444
## diabetes                                0.987897
## totChol                                 0.165826
## sysBP                                   0.000462 ***
## diaBP                                   0.341558
## BMI                                     0.773177
## heartRate                               0.999866
## glucose                                 0.002132 **
## ---
```

```
## 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: 1930.7  on 2543  degrees of freedom
## AIC: 1966.7
##
## Number of Fisher Scoring iterations: 5
```

The formula for this logistic model is as follows:

Chance 10 year CHD = 1/ (1+ e^-(-8.495 + B1(Whether or not you are male) + B2(Age) + B3(Do you have a high school education)+ B4(Do you have some college education)+ B5(do you have some high school education)+ B6(are you a smoker)+ B7(How many cigarettes per day do you smoke)+B8(are you on BP meds)+ B9(Previous stroke)+B10(currently hypertensive)+ B11(Currently has diabetes)+ B12(total cholesterol)+B13(systolic blood pressure)+ B14(Diastolic bp)+ B15(BMI)+ B16(heart rate)+ B17(glucose)))

B1 = .43; B2=.065; B3=-.12; B4=-.077; B5=.064; B6=.103; B7=.0174; B8=-.107; B9=.937; B10=.244; B11=-.0059; B12=.00187; B13=.0168; B14=-.007463; B15=.0045; B16=-.00000084; B17=.0084 Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Problem 3 part a ii

Based on p-values, the only factors which are significant are gender, age, cigarettes per day, systolic blood pressure, and blood glucose. The coefficient of blood glucose is positive, meaning that higher blood glucose levels at time of evaluation are related to a greater risk of coronary heart disease within the next 10 years.Unlike in the case of linear regression, the change in probability of CHD and the change in blood glucose levels are not linearly related. (maybe add more to this) ### Problem 3 part a iii

Setting the cost of medicating equal to not medicating: 500,000p = (1-p/4)60,000+(p/4)560,000

The break even point for p is .16

Problem 3 part a iv

Confusion matrix using the threshold of .16

```
chd.test_predTest = predict(mod1, newdata=chd.test, type="response")

summary(chd.test_predTest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01573 0.06554 0.11854 0.15780 0.20564 0.93914
```

```
table(chd.test$TenYearCHD, chd.test_predTest > 0.16)
```

```
##
##     FALSE TRUE
##   0   637  293
##   1    56  111
```

```
table(chd.test$TenYearCHD)
```

```
##
##   0   1
## 930 167
```

```
accuracy = (637+111)/(1097)
TPR = 111/(56+111)
FPR = 293/(293+637)
```

accuracy:

```
## [1] 0.6818596
```

TPR:

```
## [1] 0.6646707
```

FPR:

```
## [1] 0.3150538
```

Problem 3 part a v

based on the decision matrix, cost of true negative is 0, cost of false negative is 500,000, cost of true positive is 560,000, and cost of false positive is 60,000. Applying this to the numbers from the confusion matrix yields the following:

```
total_cost = 0*637+ 500000*56+60000*293+560000*111
cost_per_person = total_cost/1097
cost_per_person
```

```
## [1] 98213.31
```

Given that we are told in the setup of the problem that the chance of CHD if medicated is 1/4 the chance of CHD if unmedicated, the assumption that medication does not change outcomes and therefore should not be considered in costs is not a valid assumption. Utilizing this knowledge, the costs of not medicating remain the same, while the costs of medicating shift, as only 1/4 of the true positives would be assumed to get sick if they were medicated.

```
total_cost_med = (0*637)+(500000*56)+((111/4)*560000)+(((111*.75)+293)*60000)
cost_per_person_med = total_cost_med/1097
cost_per_person_med
```

```
## [1] 60268.92
```

Problem 3 part a vi

Simple baseline model predicting that noone should get medication:

```
table(chd.test$TenYearCHD, chd.test_predTest > 1)
```

```
##
##     FALSE
##   0   930
##   1   167
```

```
accuracy_baseline = 930/1097
accuracy_baseline
```

```
## [1] 0.8477666
```

```
cost_per_person_baseline = (167*500000)/1097
cost_per_person_baseline
```

```
## [1] 76116.68
```

True positive rate is 0 because the baseline model predicts that noone will get CHD. False positive rate is also zero, because once again the baseline model predicts that noone will get CHD. The accuracy of the baseline model is .85, which is higher than the accuracy of the applied model- however it is important to note that the cost of a false negative is far higher than the cost of a false positive, so decision making based on the model saves money (~$16,000 per person). The cost per person is $76,117

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.

Problem B part a vii

Predicting CHD rate for a new 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.

```
new_patient <- data.frame(male=0, age = 51, education = "College", currentSmoker = 1,
cigsPerDay = 20, BPMeds = 0, prevalentStroke = 0, prevalentHyp = 1, diabetes = 0, totC
hol = 220, sysBP = 140, diaBP = 100, BMI = 31, heartRate = 59, glucose = 78)

predict(mod1, newdata=new_patient, type="response")
```

```
##         1
## 0.1567618
```

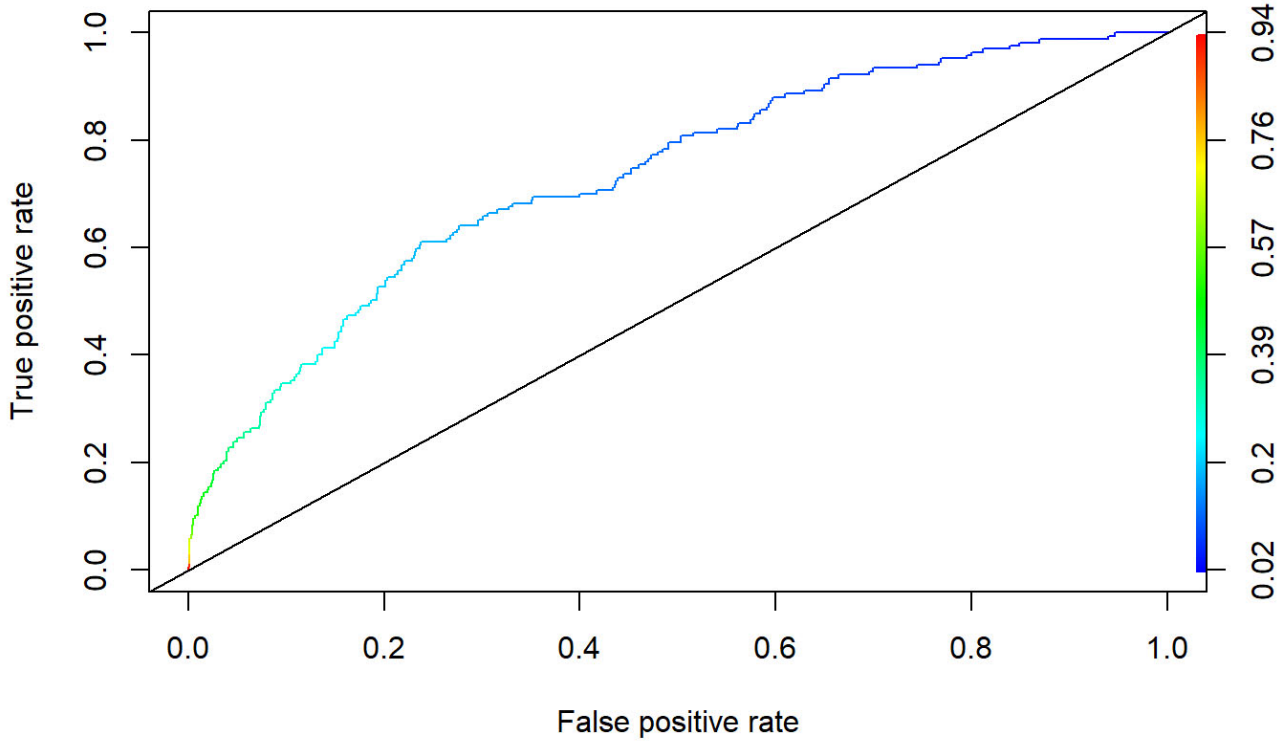
The model predicts that this patient has an unmedicated risk of .1567618, meaning she falls just short of the .16 threshold and should not be medicated.

Problem 3 part b

ROC curve for the test set and auc calculations described below. The AUC for this model on this test set is .7335716. ROC curves can ve very helpful in determining either improvements in the modeling upon which decisions surrounding the adoption of a particular medication option or alternatives between different medications. In all cases, a higher ROC indicates lower rates of false positives relative to true positives, meaning that the model is performing better.

One interesting aspect of this ROC curve is a plateau near the middle- beginning at a false positive rate of around .39, there is lettle to no improvement in the true positive rate until the false positive rate reaches nearly .45.

```
rocr.log.pred <- prediction(chd.test_predTest, 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)
```

```
## [1] 0.7335716
```

Problem 3 part c

Decision tree for customers reaches the following (assuming probability = p = .16 and copay = C)

Decision to copay => get CHP ((C+300000)* p/4) or no CHP ((C)*(1-p/4)) Decision to not copay => get CHP (300000* p) or no CHP (0 * (1-p))

Set the decision to copay and decision not to copay equal, set probability to .16, solve for C:

the copay would need to be $36,000 in order to incentivize customers to self select such that they will only choose to go on medication if they have a greater than 16% chance of getting CHD.

Problem 3 part d

One of the major ethical issues thusfar unaddressed in this analysis is the fact that the only value placed on human life by the insurance company is the cost of healthcare for those who get sick. One way of combatting this would be to add a cost factor, possibly by increasing the percieved ‘cost’ of

people getting CHD beyond the costs payed by the healthcare system to account for the value of patient health and happiness. Another ethical issue in the analysis so far is that it ignores human psychology and loss aversion- Even though the rational patient decision tree dictates that it would be most logical for patients to elect to pay the high copay to receive treatment if they have a greater than .16 risk of CHD, in reality patients with much higher risks would likely forego treatment because the short term chosen cost of the high copay is much more concrete and easier to wrap ones head around than the long-term potential cost of getting the disease.