Tim Miller
15.071 | HW #2

Problem 1a

```
> mod <- glm(data=train, TenYearCHD ~ ., family='binomial')
> summary(mod)

Call:
glm(formula = TenYearCHD ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7949  -0.5989  -0.4271  -0.2920   2.7796

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -8.0970009  0.8426453  -9.609  < 2e-16 ***
male1                                  0.3792662  0.1301201   2.915  0.00356 **
age                                    0.0600884  0.0081157   7.404 1.32e-13 ***
educationHigh school/GED               0.0190142  0.2215188   0.086  0.93160
educationSome college/vocational school  0.1230268  0.2398872   0.513  0.60805
educationSome high school              0.2083648  0.2074184   1.005  0.31511
currentSmoker1                         0.2425214  0.1870299   1.297  0.19473
cigsPerDay                             0.0155637  0.0075782   2.054  0.04000 *
BPMeds1                                0.4618355  0.2688809   1.718  0.08587 .
prevalentStroke1                       0.8750945  0.5677020   1.541  0.12320
prevalentHyp1                          0.3785068  0.1646511   2.299  0.02151 *
diabetes1                             -0.0452063  0.3823421  -0.118  0.90588
totChol                                0.0031881  0.0013626   2.340  0.01929 *
sysBP                                  0.0108658  0.0048251   2.252  0.02433 *
diaBP                                 -0.0006302  0.0079933  -0.079  0.93716
BMI                                    0.0005710  0.0149748   0.038  0.96958
heartRate                             -0.0036763  0.0049755  -0.739  0.45997
glucose                                0.0071727  0.0026856   2.671  0.00757 **
---
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: 1935.8  on 2543  degrees of freedom
AIC: 1971.8

Number of Fisher Scoring iterations: 5
```

Based on the model where nothing was excluded, the following risk factors are significant:

- If the patient is a male i.e., 'male1'
- Age
- cigsPerDay
- prevalentHyp
- totChol
- sysBP – Systolic blood pressure
- glucose – blood glucose level

Intuitively all these variables make sense, especially cigsPerDay, sysBP, and prevalentHP. Intuitively I know that smokers, people with high blood pressure, and have hypertension all are at risk for heart disease. Interested to see that being a current smoker is not significant. Perhaps this is an indication that people who occasionally smoke are not as at risk.

Problem 1b

```
> predict(mod, newdata=fourth.patient.copies, type = "response")
        4       4.1
0.3971968 0.3408114
```

The probability of 10 Year CHD if the patient was a non-smoker is ~34.08% vs. a higher 39.72% for the patient's current condition as a smoker.

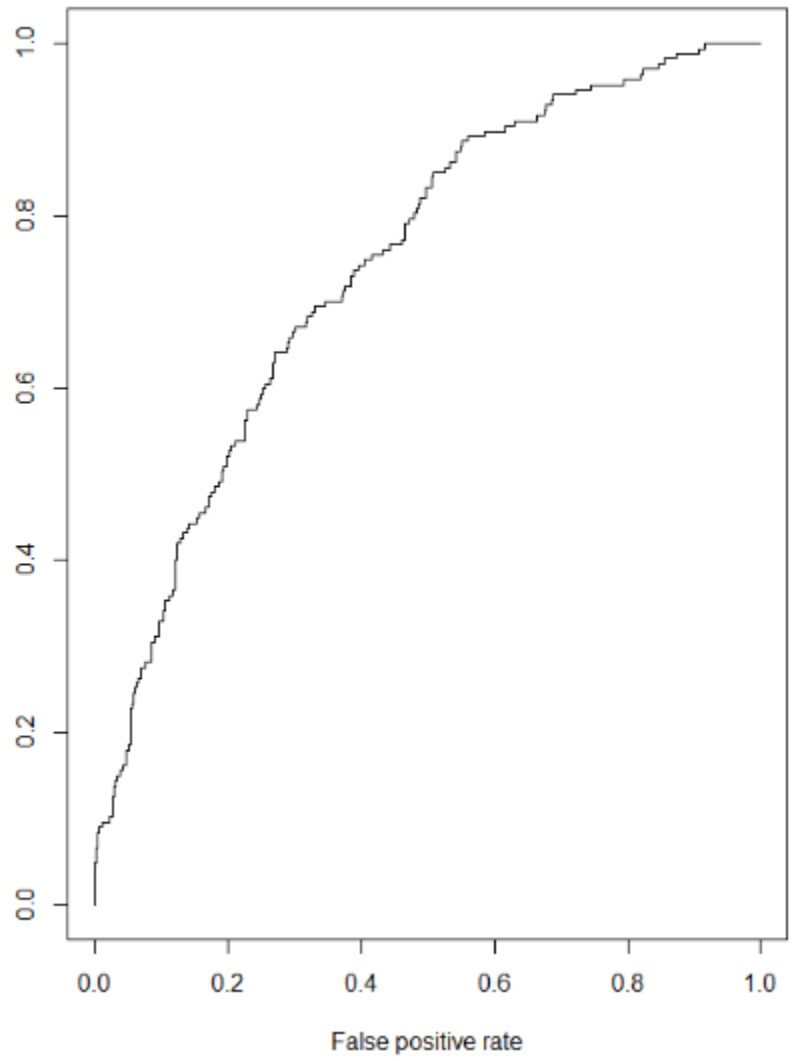
Problem 1c

I do not agree with this statement. There are numerous other factors that contribute to 10-year CHD. Just stopping smoking does not correct for those other factors. For example, we do not know how much of an effect prior smoking habits had on the person's blood pressure. Perhaps they were a very heavy smoker and really affected their blood pressure past the point of no return. The rapid rise in blood pressure from past smoking could outweigh the potential benefits gained from stopping smoking.

Problem 1d

Test set ROC

```
> pred <- predict(mod, test, type = "response")
> rocr.pred = prediction(pred, test$TenYearCHD)
> plot(performance(rocr.pred, "tpr", "fpr"))
>
```



AUC

```
> AUC = as.numeric(performance(rocr.pred, "auc")@y.values)
> AUC
[1] 0.7406542
>
```

The area under the curve is ~0.741.

Problem 2a

*Patient that does **not** take medicine:*
Expected cost = 190,000p + (1 – p)* 0 = 190,000p

*Patient that **does** take medicine:*
Expected cost = 200,000(p/3) + (1 – p/3)*10,000

To solve for p, we need to set the two equations equal
190,000p = 200,000(p/3) + (1 – p/3)*10,000
Rearranging and solving for p, we get **p = 0.078947 or ~7.8947%**

For any values of p greater than 7.8947%, I would recommend medication. This is when expected cost of medication is less than expected cost of no medication.

Problem 2b

Confusion matrix

```
# Create new prediction column
test$probs = predict(mod, newdata=test, type="response")

# Cutoff
cutoff = 10000 / (190000 - 200000/3 + 10000/3)
cutoff

#Re-create a "user friendly" version of the 'outcome' column
test = mutate(test, actual_outcome = ifelse(TenYearCHD == 1,"CHD","NoCHD"))

# Prediction measure
test = mutate(test, prediction = ifelse(probs >= cutoff, "Medication","NoMedication"))

# view mutated dataframe
head(test)
tail(test)

# confusion matrix
confusion_matrix = table(test$prediction, test$actual_outcome)

# Take a look
confusion_matrix
```

```
              CHD NoCHD
Medication    152   586
NoMedication   15   344
>
```

Accuracy

```
> # ACCURACY
> accuracy = (confusion_matrix[1,1] + confusion_matrix[2,2]) / nrow(test)
> accuracy
[1] 0.4521422
>
```

The model has an accuracy of ~45.21%.

TPR

```
> # TPR
> TPR = confusion_matrix[2,2] / (confusion_matrix[1,2] + confusion_matrix[2,2])
> TPR
[1] 0.3698925
```

The model has a True Positive Rate of ~36.99%.

FPR

```
> FPR = confusion_matrix[2,1] / (confusion_matrix[2,1] + confusion_matrix[2,2])
> FPR
[1] 0.04178273
```

The model has a false positive rate of ~41.78%

Problem 2c

Extremely busy week and ran out of time for this problem before submission deadline

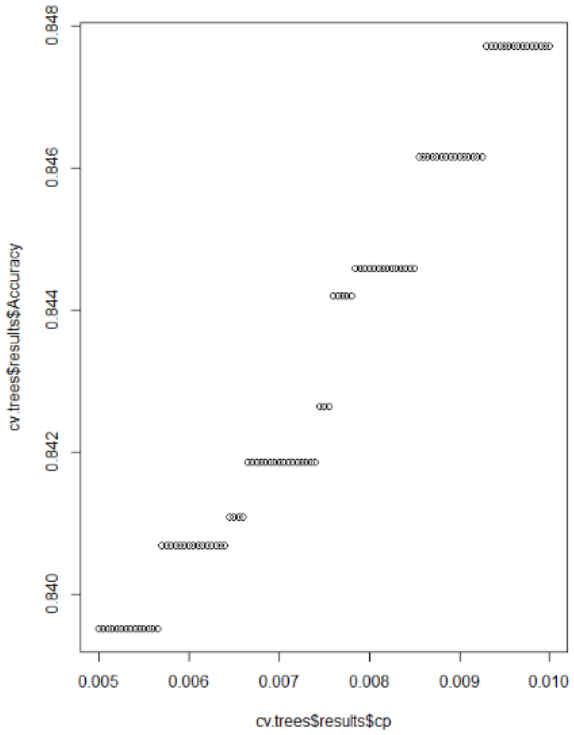
Problem 2d

Extremely busy week and ran out of time for this problem before submission deadline

Problem 2e

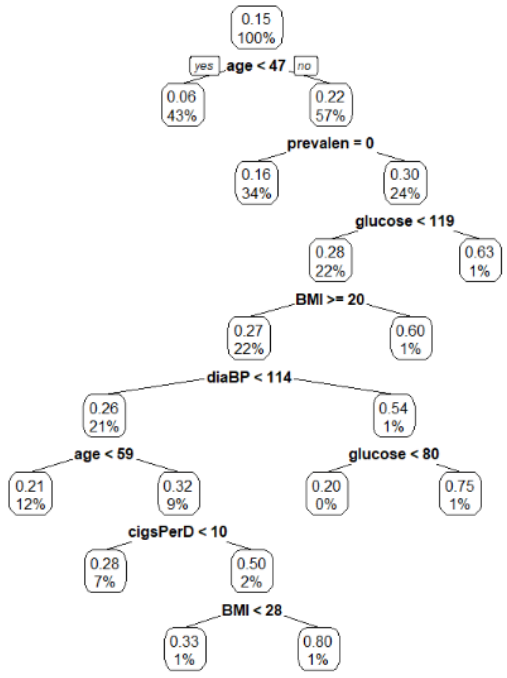
The calculation of quality of life and life expectancy reminds me of the Ford Pinto controversy. In that case and here, we are assigning a value to life quality and life expectancy. While I know we are trying to get to a numerical answer, assigning and actual value is very difficult. Some people might argue that we should not make things black and white. Instead we should give medicine to people that need it. So even if the model predicts that the person is not about the p value threshold, they should still get access to the medicine as a preventative measure.

Problem 3a



The cp parameters from 0.00925 to 0.01 all look to give the best result, but that is likely due to rounding. Therefore, we use cv.trees$bestTune to determine the best cp value here is 0.01.

Problem 3b



The group with the lowest probability of CHD is individuals under the age of 47. That includes 43% of the population.

Problem 3c

This node is saying that people with BMI higher than 28 have a higher probability of CHD than those with CHD less than 28.

This is counterintuitive because the earlier split say that people with greater than or equal to 20 BMI are at less risk of CHD than those with less than or equal to 20 BMI.

To explain this, I would say that for the very bottom split we are introducing the impacts of smoking, which we know from 1a are significant. So when someone is a smoker of >10 packs a day, the effect of higher BMI is more significant than if the person is not smoker.

Problem 3d

Based on my answer to Problem 3c, I do not think logistic regression correctly captures the nuanced relationship of how smoking affects BMI relationship with CHD. Therefore it is probably not the best model to use.

Problem 3e

Extremely busy week and ran out of time for this problem before submission deadline

Problem 3f

If I did not run out of time, I would have compared the ROC and AUC between the two models. My assumption is that they are pretty close therefore equal in prediction power. Since the tree model is much easier to visually look at and iterate through I would say the tree model is more practical. With logistic regression, non technical people might have a hard time understanding it.