Modern Data Mining, HW 4

Group Member Hannah Xiao Group Member Kateryna Suprun
Group Member Kyle Liao

11:59 pm, 03/19, 2023

Contents

1 Overview

Logistic regression is used for modeling categorical response variables. The simplest scenario is how to identify risk factors of heart disease? In this case the response takes a possible value of `YES` or `NO`. Logit link function is used to connect the probability of one being a heart disease with other potential risk factors such as `blood pressure`, `cholestrol level`, `weight`. Maximum likelihood function is used to estimate unknown parameters. Inference is made based on the properties of MLE. We use AIC to help nailing down a useful final model. Predictions in categorical response case is also termed as `Classification` problems. One immediately application of logistic regression is to provide a simple yet powerful classification boundaries. Various metrics/criteria are proposed to evaluate the quality of a classification rule such as `False Positive`, `FDR` or `Mis-Classification Errors`.

LASSO with logistic regression is a powerful tool to get dimension reduction.

1.1 Objectives

- Understand the model
  - logit function
    - interpretation
  - Likelihood function
- Methods
  - Maximum likelihood estimators
    - Z-intervals/tests
    - Chi-squared likelihood ratio tests
- Metrics/criteria
  - Sensitivity/False Positive
  - True Positive Prediction/FDR
  - Misclassification Error/Weighted MCE
  - Residual deviance
  - Training/Testing errors
- LASSO
- R functions/Packages
  - `glm()`, `Anova`
  - `pROC`
  - `cv.glmnet`

1.2 Review

Review the code and concepts covered in

- Module Logistic Regressions/Classification
- Module LASSO in Logistic Regression

1.3 This homework

We have two parts in this homework. Part I is guided portion of work, designed to get familiar with elements of logistic regressions/classification. Part II, we bring you projects. You have options to choose one topic among either Credit Risk via LendingClub or Diabetes and Health Management. Find details in the projects.

2 Part I: Framingham heart disease study

We will continue to use the Framingham Data (`Framingham.dat`) so that you are already familiar with the data and the variables. All the results are obtained through training data.

Liz is a patient with the following readings: `AGE=50, GENDER=FEMALE, SBP=110, DBP=80, CHOL=180, FRW=105, CIG=0`. We would be interested to predict Liz's outcome in heart disease.

To keep our answers consistent, use a subset of the data, and exclude anyone with a missing entry. For your convenience, we've loaded it here together with a brief summary about the data.

We note that this dataset contains 311 people diagnosed with heart disease and 1095 without heart disease.

After a quick cleaning up here is a summary about the data:

```
   HD             AGE           SEX           SBP            DBP
   0:1086   Min.   :45.0   FEMALE:730   Min.   : 90   Min.   : 50.0
   1: 307   1st Qu.:48.0   MALE  :663   1st Qu.:130   1st Qu.: 80.0
            Median :52.0                Median :142   Median : 90.0
            Mean   :52.4                Mean   :148   Mean   : 90.2
            3rd Qu.:56.0                3rd Qu.:160   3rd Qu.: 98.0
            Max.   :62.0                Max.   :300   Max.   :160.0
        CHOL           FRW            CIG
   Min.   : 96   Min.   : 52   Min.   : 0
   1st Qu.:200   1st Qu.: 94   1st Qu.: 0
   Median :230   Median :103   Median : 0
   Mean   :235   Mean   :105   Mean   : 8
   3rd Qu.:264   3rd Qu.:114   3rd Qu.:20
   Max.   :430   Max.   :222   Max.   :60
```

Lastly we would like to show five observations randomly chosen.

```
##     HD AGE    SEX SBP DBP CHOL FRW CIG
## 643  1  61   MALE 140  68  248 104  20
## 11   0  45   MALE 110  88  183  90   0
## 576  1  58   MALE 150  95  296 100  15
## 560  1  59   MALE 260 130  246 111  20
## 702  0  45 FEMALE 122  74  178  88   5
```

2.1 Identify risk factors

2.1.1 Understand the likelihood function

Conceptual questions to understand the building blocks of logistic regression. All the codes in this part should be hidden. We will use a small subset to run a logistic regression of `HD` vs. `SBP`.

i. Take a random subsample of size 5 from `hd_data_f` which only includes `HD` and `SBP`. Also set `set.seed(471)`. List the five observations neatly below. No code should be shown here.

```
##     HD SBP
## 643  1 140
## 11   0 110
## 576  1 150
## 560  1 260
## 702  0 122
```

ii. Write down the likelihood function using the five observations above. L = exp(B0+B1*140)/(1+exp(B0+B1*140)) + 1/(1+exp(B0+B1*110)) + exp(B0+B1*150)/(1+exp(B0+B1*150)) + exp(B0+B1*260)/(1+exp(B0+B1*260)) + 1/(1+exp(B0+B1*122))

iii. Find the MLE based on this subset using glm(). Report the estimated logit function of `SBP` and the probability of `HD`=1. Briefly explain how the MLE are obtained based on ii. above. The logit function is: logit = -334.96 + 2.56SBP, and the probability of HD = 1 os P(HD=1|SB) = $e^{\text{logit}/(1+e}$logit), where logit is given earlier. The MLE is obtained by maximizing the likelihood function. We first take the log of the likelihood function because the maximum of a function is the same as the maximum of the log of a function. We then take the derivative of that with respect to the coefficient of each predictor variable and set it equal to 0 to find the coefficient that maximizes the likelihood function. Repeating this for each coefficient gives the maximum likelihood estimate. The MLE for beta0 is -334.96 and the MLE for beta1 is 2.56

iv. Evaluate the probability of Liz having heart disease.

```
##        1
## 2.22e-16
```

2.1.2 Identify important risk factors for `Heart.Disease`.

We focus on understanding the elements of basic inference method in this part. Let us start a fit with just one factor, `SBP`, and call it `fit1`. We then add one variable to this at a time from among the rest of the variables. For example

```
##
## Call:
## glm(formula = HD ~ SBP, family = binomial, data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.661  -0.709  -0.624  -0.524   2.107
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.65489    0.34787  -10.51  < 2e-16 ***
## SBP          0.01581    0.00222    7.12  1.1e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1469.3  on 1392  degrees of freedom
```

```
## Residual deviance: 1417.5  on 1391  degrees of freedom
## AIC: 1421
##
## Number of Fisher Scoring iterations: 4

##
## Call:
## glm(formula = HD ~ SBP + AGE, family = binomial, data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.630  -0.721  -0.601  -0.466   2.169
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.48625    0.79182   -8.19  2.6e-16 ***
## SBP          0.01434    0.00225    6.38  1.8e-10 ***
## AGE          0.05775    0.01422    4.06  4.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1469.3  on 1392  degrees of freedom
## Residual deviance: 1400.8  on 1390  degrees of freedom
## AIC: 1407
##
## Number of Fisher Scoring iterations: 4

##
## Call:
## glm(formula = HD ~ SBP + SEX, family = binomial, data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.641  -0.737  -0.573  -0.417   2.245
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.57026    0.38973  -11.73  < 2e-16 ***
## SBP          0.01872    0.00232    8.05  8.1e-16 ***
## SEXMALE      0.90342    0.13976    6.46  1.0e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1469.3  on 1392  degrees of freedom
## Residual deviance: 1373.8  on 1390  degrees of freedom
## AIC: 1380
##
## Number of Fisher Scoring iterations: 4

##
```

```
## Call:
## glm(formula = HD ~ SBP + DBP, family = binomial, data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.660  -0.710  -0.623  -0.522   2.096
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.75988    0.41651   -9.03  < 2e-16 ***
## SBP          0.01449    0.00364    3.98  6.8e-05 ***
## DBP          0.00334    0.00726    0.46     0.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1469.3  on 1392  degrees of freedom
## Residual deviance: 1417.3  on 1390  degrees of freedom
## AIC: 1423
##
## Number of Fisher Scoring iterations: 4

##
## Call:
## glm(formula = HD ~ SBP + CHOL, family = binomial, data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.734  -0.714  -0.622  -0.507   2.159
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.19172    0.46354   -9.04  < 2e-16 ***
## SBP          0.01539    0.00224    6.88  5.9e-12 ***
## CHOL         0.00254    0.00142    1.79    0.074 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1469.3  on 1392  degrees of freedom
## Residual deviance: 1414.3  on 1390  degrees of freedom
## AIC: 1420
##
## Number of Fisher Scoring iterations: 4

##
## Call:
## glm(formula = HD ~ SBP + FRW, family = binomial, data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.677  -0.710  -0.624  -0.523   2.110
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.74427    0.44809   -8.36  < 2e-16 ***
## SBP          0.01556    0.00236    6.61  3.9e-11 ***
## FRW          0.00120    0.00377    0.32     0.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1469.3  on 1392  degrees of freedom
## Residual deviance: 1417.4  on 1390  degrees of freedom
## AIC: 1423
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = HD ~ SBP + CIG, family = binomial, data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.016  -0.709  -0.604  -0.488   2.120
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.99037    0.36607  -10.90  < 2e-16 ***
## SBP          0.01687    0.00227    7.42  1.1e-13 ***
## CIG          0.02049    0.00545    3.76  0.00017 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1469.3  on 1392  degrees of freedom
## Residual deviance: 1403.8  on 1390  degrees of freedom
## AIC: 1410
##
## Number of Fisher Scoring iterations: 4
```

i. Which single variable would be the most important to add? Add it to your model, and call the new fit `fit2`.
Sex would be the most important variable to add, as it has the next lowest p-value.

```
##
## Call:
## glm(formula = HD ~ SBP + SEX, family = binomial, data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.641  -0.737  -0.573  -0.417   2.245
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.57026    0.38973  -11.73  < 2e-16 ***
## SBP          0.01872    0.00232    8.05  8.1e-16 ***
## SEXMALE      0.90342    0.13976    6.46  1.0e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1469.3  on 1392  degrees of freedom
## Residual deviance: 1373.8  on 1390  degrees of freedom
## AIC: 1380
##
## Number of Fisher Scoring iterations: 4
```

We will pick up the variable either with highest $|z|$ value, or smallest p value. Report the summary of your `fit2` Note: One way to keep your output neat, we will suggest you using `xtable`. And here is the summary report looks like.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.5703	0.3897	-11.73	0.0000
SBP	0.0187	0.0023	8.05	0.0000
SEXMALE	0.9034	0.1398	6.46	0.0000

ii. Is the residual deviance of `fit2` always smaller than that of `fit1`? Why or why not? Residual deviance of a fit with more variables will always be smaller than that of one with fewer. This is because if the variable increases the fitting ability of the model, the residual deviance will decrease. If adding the variable to the model does not increase the fitting ability, the residual deviance will be minimally changed but in the downwards direction.

iii. Perform both the Wald test and the Likelihood ratio tests (Chi-Squared) to see if the added variable is significant at the .01 level. What are the p-values from each test? Are they the same?

The p-value from the likelihood ratio test is 3.8e-11 and the p-value from the Wald test is 1.4e-10. They are not the same. Both are less than the significance level of .01, so we reject the null that the full (SBP and Sex) model and the reduced (SBP only) model fit equally well. We should use the model with both SBP and Sex.

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Likelihood ratio test
##
## Model 1: HD ~ SBP + SEX
## Model 2: HD ~ SBP
```

```
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1   3   -687
## 2   2   -709 -1  43.7    3.8e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Wald test
##
## Model 1: HD ~ SBP + SEX
## Model 2: HD ~ SBP
##   Res.Df Df    F  Pr(>F)
## 1   1390
## 2   1391 -1 41.8 1.4e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.1.3 Model building

Start with all variables. Our goal is to fit a well-fitting model, that is still small and easy to interpret (parsimonious).

i. Use backward selection method. Only keep variables whose coefficients are significantly different from 0 at .05 level. Kick out the variable with the largest p-value first, and then re-fit the model to see if there are other variables you want to kick out. Highest p-values in order were DBP (.70594), FRW (0.1315) and CIG(0.0608), leaving SBP, AGE, SEX, and CHOL as significant.

```
##
## Call:
## glm(formula = HD ~ SBP + AGE + SEX + CHOL, family = binomial,
##     data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.607  -0.735  -0.552  -0.348   2.434
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.40872    0.90860   -9.25  < 2e-16 ***
## SBP          0.01696    0.00236    7.18  7.0e-13 ***
## AGE          0.05664    0.01450    3.91  9.4e-05 ***
## SEXMALE      0.98987    0.14505    6.82  8.8e-12 ***
## CHOL         0.00448    0.00150    3.00   0.0027 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1469.3  on 1392  degrees of freedom
## Residual deviance: 1349.0  on 1388  degrees of freedom
## AIC: 1359
##
## Number of Fisher Scoring iterations: 4
```

ii. Use AIC as the criterion for model selection. Find a model with small AIC through exhaustive search. Does exhaustive search guarantee that the p-values for all the remaining variables are less than .05? Is our final model here the same as the model from backwards elimination? Exhaustive search with AIC does not guarentee that all p-values are less than .05, and because of that, this model contains variables that were discarded in the previous model such as CIG and FRW.

```
## Morgan-Tatar search since family is non-gaussian.

##
## Call:
## glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.707  -0.728  -0.552  -0.334   2.450
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.22786    0.99615   -9.26  < 2e-16 ***
## AGE          0.06153    0.01478    4.16  3.1e-05 ***
## SEXMALE      0.91127    0.15712    5.80  6.6e-09 ***
## SBP          0.01597    0.00249    6.42  1.4e-10 ***
## CHOL         0.00449    0.00150    2.99   0.0028 **
## FRW          0.00604    0.00400    1.51   0.1315
## CIG          0.01228    0.00609    2.02   0.0437 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1469.3  on 1392  degrees of freedom
## Residual deviance: 1343.3  on 1386  degrees of freedom
## AIC: 1357
##
## Number of Fisher Scoring iterations: 4
```

iii. Use the model chosen from part ii. as the final model. Write a brief summary to describe important factors relating to Heart Diseases (i.e. the relationships between those variables in the model and heart disease). Give a definition of "important factors".

Important factors are those defined as ones that lowered increased the fitting ability of the above model to the data given. The important factors are age, sex, SBP, cholesterol levels, FRW, and cigarettes smoked. Being older, male, and having higher SBP, CHOL, FRW and smoking more cigarettes increases the probability of having heart diesease as they decrease the logit.

iv. What is the probability that Liz will have heart disease, according to our final model? Liz has a probability of .0496 of having heart disease.

```
##      1
## 0.0496
```

2.2 Classification analysis

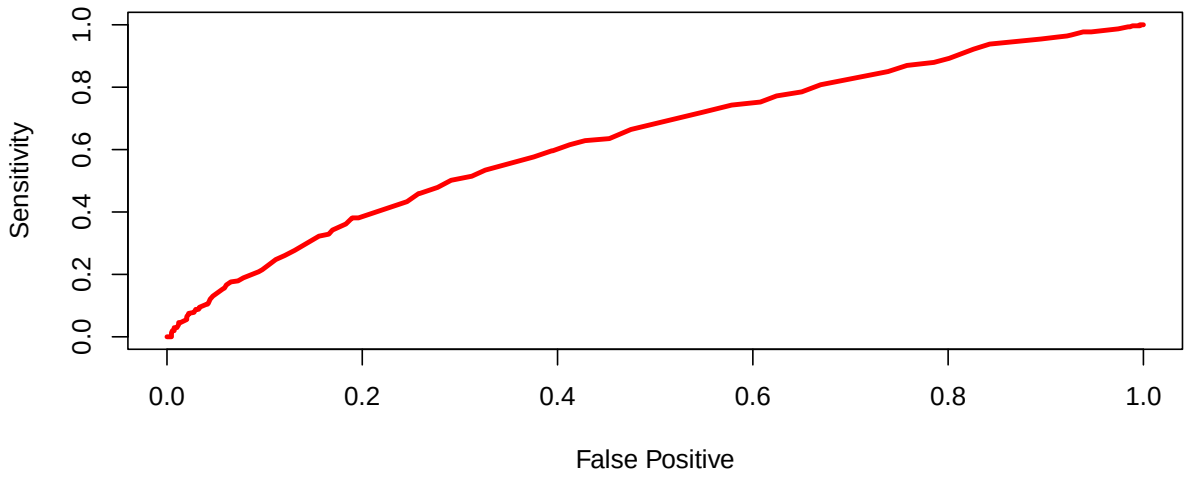
2.2.1 ROC/FDR

i. Display the ROC curve using `fit1`. Explain what ROC reports and how to use the graph. Specify the classifier such that the False Positive rate is less than .1 and the True Positive rate is as high as possible.

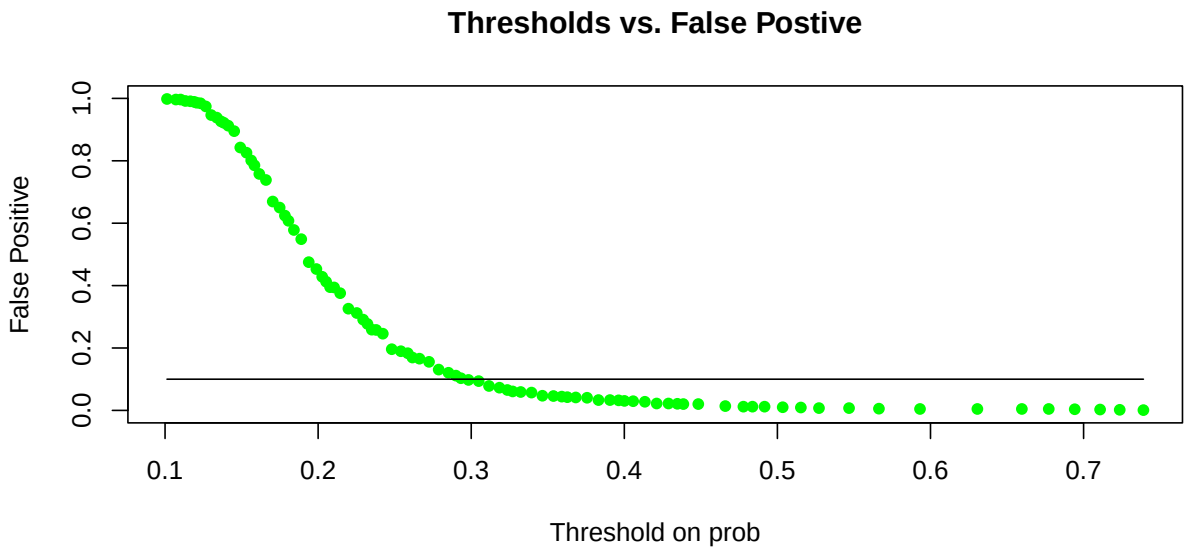
The ROC curve is a graph with the False Positive Rate on the x-axis and the True Positive Rate (Sensitivity) on the y-axis as the threshold changes.

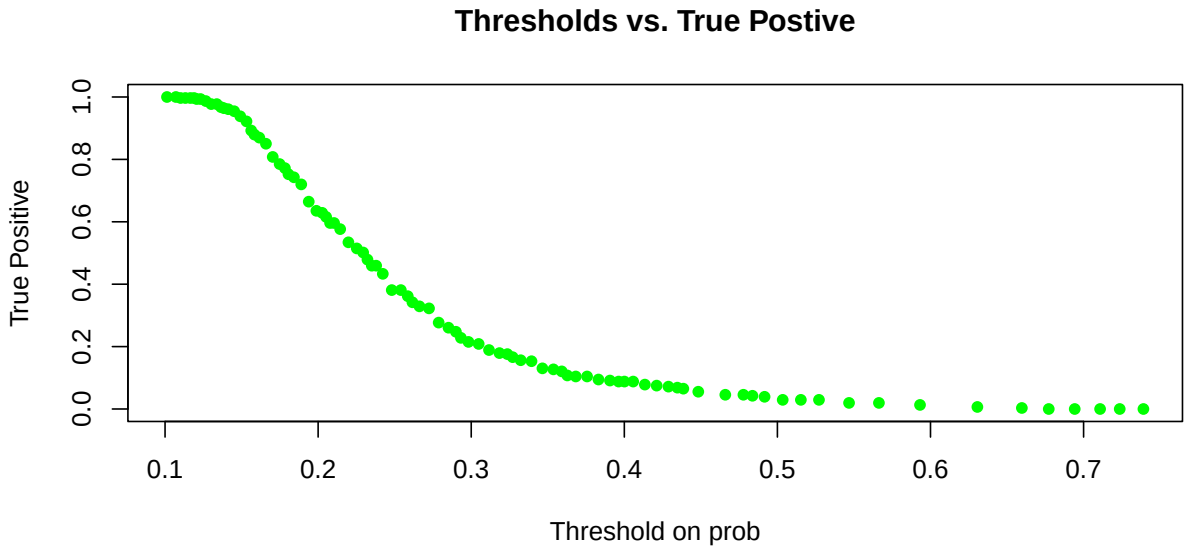
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



Because True positive is always decreasing, maximizing the true positive rate means we have to pick the threshold with the highest possible false posivie rate. Picking the threshold where the false positive rate is at a maximum at .1 yields a threshold of around .3.

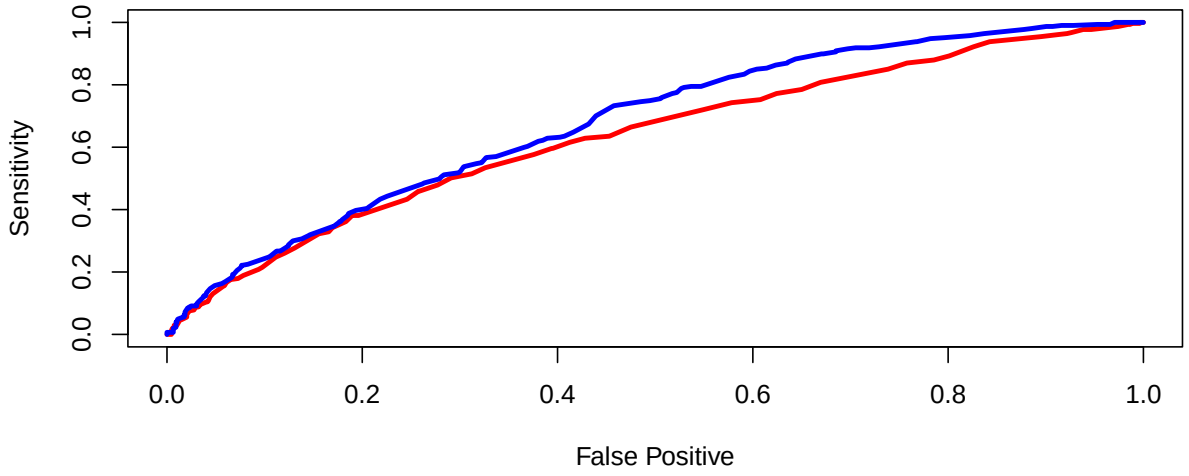




ii. Overlay two ROC curves: one from `fit1`, the other from `fit2`. Does one curve always contain the other curve? Is the AUC of one curve always larger than the AUC of the other one? Why or why not? One curve does not always have to contain or be larger than the other. Though generally fit 2 is outside of fit 1, when False Positive rate is low, there are some thresholds that make fit 1 outside of fit 2. This is because some classifiers may be more accurate at lower thresholds and less accurate at higher thresholds, or vice versa, which would make them intersect.

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

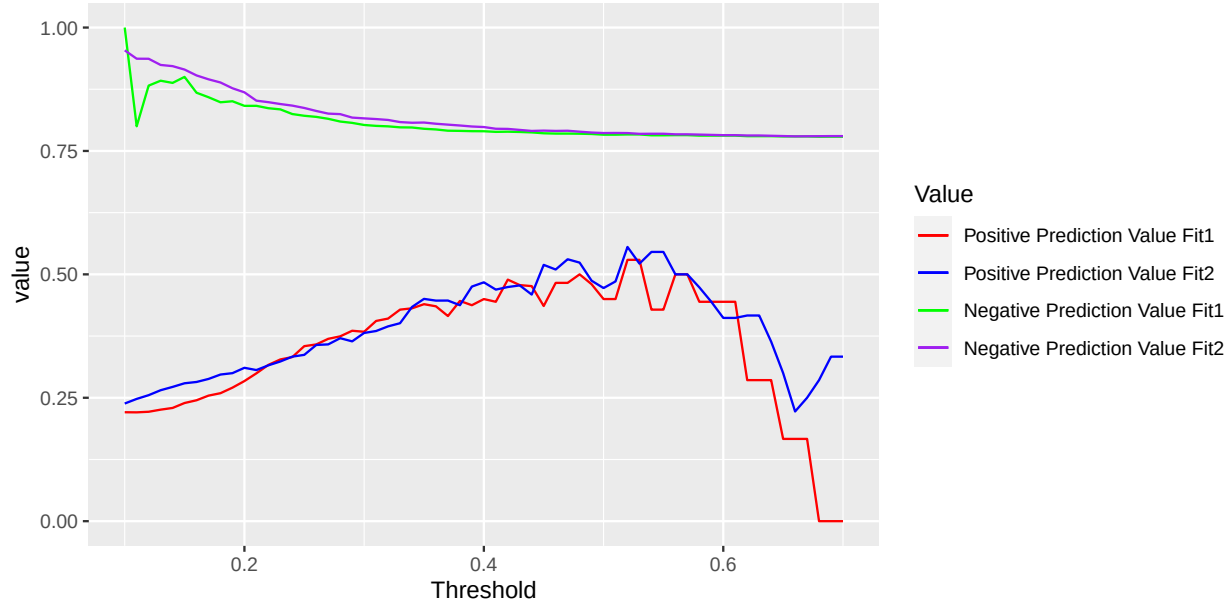


iii. Estimate the Positive Prediction Values and Negative Prediction Values for `fit1` and `fit2` using .5 as a threshold. Which model is more desirable if we prioritize the Positive Prediction values? If Positive Prediction value is more desirable, we would rather use fit2 because it has a larger Positive Prediction value.

```
## [1] "Positive Prediction Value for fit 1 is  0.45 , and negative prediction value is  0.7829570284049
```

```
## [1] "Positive Prediction Value for fit 2 is  0.472222222222222 , and negative prediction value is  0
```

iv. For `fit1`: overlay two curves, but put the threshold over the probability function as the x-axis and positive prediction values and the negative prediction values as the y-axis. Overlay the same plot for `fit2`. Which model would you choose if the set of positive and negative prediction values are the concerns? If you can find an R package to do so, you may use it directly.



```
## numeric(0)
```

```
## [1] 0.52
```

2.2.2 Cost function/ Bayes Rule

Bayes rules with risk ratio $\frac{a_{10}}{a_{01}} = 10$ or $\frac{a_{10}}{a_{01}} = 1$. Use your final model obtained from Part 1 to build a class of linear classifiers.

i. Write down the linear boundary for the Bayes classifier if the risk ratio of $a_{10}/a_{01} = 10$. P(Y=1|x) > .1/(1+.1) = .0909

logit = -9.228 + .06153AGE + .91127SEX + .01597SBP + .00449CHOL + .00604FRW + .01228CIG > log(.0909/.9090) = -2.3026

ii. What is your estimated weighted misclassification error for this given risk ratio?

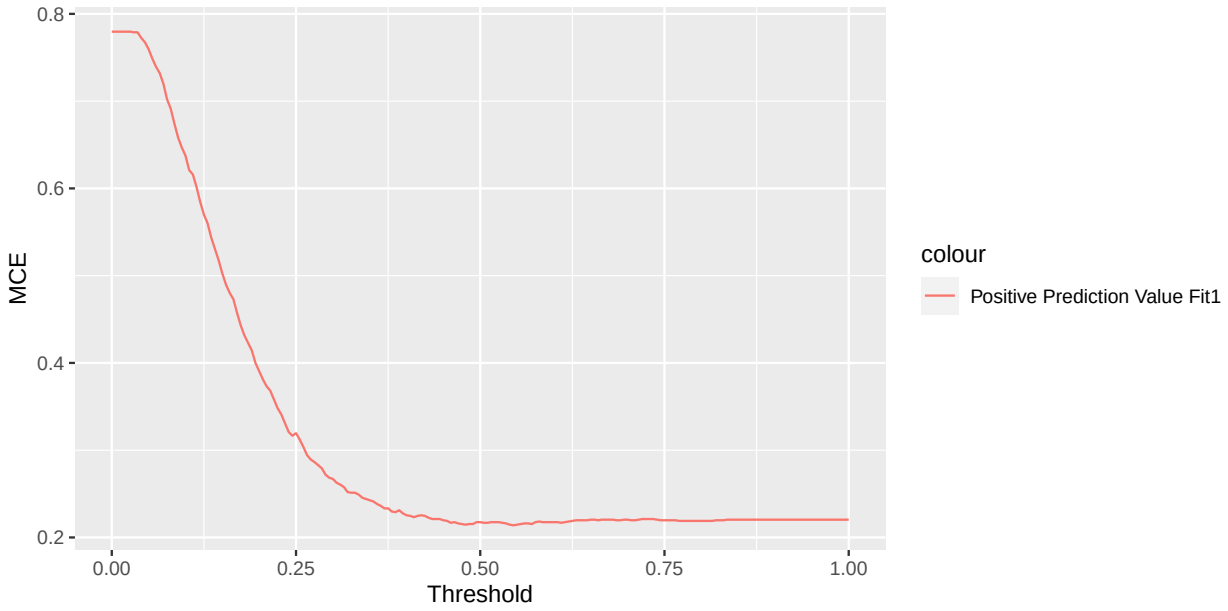
```
## [1] 0.714
```

iii. How would you classify Liz under this classifier? She would not have heart disease.

```
##   1
## "0"
```

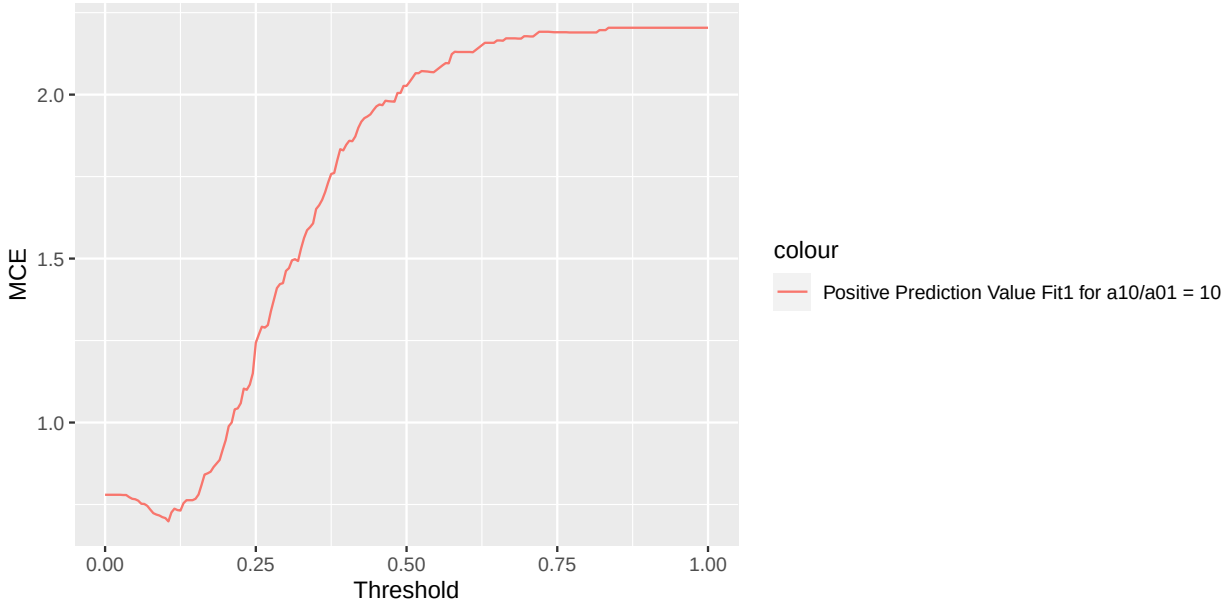
iv. Bayes rule gives us the best rule if we can estimate the probability of `HD-1` accurately. In practice we use logistic regression as our working model. How well does the Bayes rule work in practice? We hope to show in this example it works pretty well.

Now, draw two estimated curves where x = threshold, and y = misclassification errors, corresponding to the thresholding rule given in x-axis.



v. Use weighted misclassification error, and set $a_{10}/a_{01} = 10$. How well does the Bayes rule classifier

perform? The Bayes rule classifer does not reach a low classification error at a relatively low threshold, showing that it is a not good classifier.



vi. Use weighted misclassification error, and set $a_{10}/a_{01} = 1$. How well does the Bayes rule classifier perform? The Bayes rule classifer does reach a low classification error at a relatively low threshold, showing that it is a good classifier.