Modern Data Mining, HW 4

Alice Lepique Fabio Oliveira Jimmy Ren

11:59 pm, 03/20, 2021

Contents

1 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:

1.1 Identify risk factors

1.1.1 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(50)`. List the five observations neatly below. No code should be shown here.

```
##      HD SBP
## 1392  1 152
## 11    0 110
## 820   0 154
## 1119  1 160
## 863   0 182
```

ii. Write down the likelihood function using the five observations above.

- L(B0, B1 | Data) = Prob(HD=1|SBP=152) x Prob(HD=0|SBP=110) x Prob(HD=0|SBP=154) x Prob(HD=1|SBP=160) x Prob(HD=0|SBP=182)
- L(B0, B1) = ($e^{(B0+152xB1))/(1+e}$(B0+152xB1)) + 1 /(1+eˆ(B0+110xB1))+ 1 /(1+eˆ(B0+154*B1)) + ($e^{(B0+160xB1))/(1+e}$(B0+160xB1)) + ($e^{(B0+182xB1))/(1+e}$(B0+182xB1))

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 MLE, or Maximum Likelihood Estimators, is the estimate that maximizes the likelihood function written above. Through glm it is obtained equivalently by minimizing the negative log of the likelihood function above.

As per the summary below, it follows:

- Logit function: logit(P(HD=1|SBP)) = -2.5456 + 0.0140 * SBP
- Probability of HD=1: P(HD=1|SBP) = (eˆ(-2.5456+0.0140 * SBP))/(1+eˆ(-2.5456+0.0140 * SBP))

```
##
## Call:
## glm(formula = HD ~ SBP, family = binomial(logit), data = hd_data_sub)
##
## Deviance Residuals:
##   1392      11     820    1119     863
##  1.357  -0.791  -1.019   1.307  -1.181
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5456     6.4069   -0.40     0.69
## SBP           0.0140     0.0413    0.34     0.73
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.7301  on 4  degrees of freedom
## Residual deviance: 6.6089  on 3  degrees of freedom
## AIC: 10.61
##
## Number of Fisher Scoring iterations: 4
```

iv. Evaluate the probability of Liz having heart disease.

The probability of Liz having heart disease is 24.2% given her SBP.

1.1.2 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. Below, the summary and Anova of the final iteration:

```
##
## Call:
## glm(formula = HD ~ SBP + AGE + SEX + DBP + CHOL + FRW + CIG,
##     family = binomial, data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.705  -0.727  -0.556  -0.333   2.446
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.33480    1.03663   -9.00  < 2e-16 ***
## SBP          0.01484    0.00389    3.82  0.00013 ***
## AGE          0.06249    0.01500    4.17  3.1e-05 ***
## SEXMALE      0.90610    0.15764    5.75  9.0e-09 ***
## DBP          0.00288    0.00762    0.38  0.70594
## CHOL         0.00446    0.00151    2.96  0.00305 **
## FRW          0.00580    0.00406    1.43  0.15296
## CIG          0.01231    0.00609    2.02  0.04315 *
## ---
## 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: 1343.1  on 1385  degrees of freedom
## AIC: 1359
##
## Number of Fisher Scoring iterations: 4

## Analysis of Deviance Table (Type II tests)
##
## Response: HD
##      LR Chisq Df Pr(>Chisq)
## SBP      14.7  1    0.00013 ***
## AGE      17.6  1    2.7e-05 ***
## SEX      34.0  1    5.4e-09 ***
## DBP       0.1  1    0.70599
## CHOL      8.7  1    0.00313 **
## FRW       2.0  1    0.15491
## CIG       4.0  1    0.04437 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i. Which single variable would be the most important to add? Add it to your model, and call the new fit `fit2`.

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.

From the summary and Anova of fit1.6, we can observe that the variable with smallest p-value is SEX, but given it is a categorical, we decided to select the next best predictor, AGE, for fit2.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.4863	0.7918	-8.19	0.0000
SBP	0.0143	0.0022	6.38	0.0000
AGE	0.0577	0.0142	4.06	0.0000

ii. Is the residual deviance of `fit2` always smaller than that of `fit1`? Why or why not?

The residual deviance of fit2 is smaller than that of fit1 (1400.8 vs 1417.5), and will always be so because the log likelihood of a bigger model in terms of number of variables will always be larger.

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?

We observe from the tests below that the variable s significant at the 0.01 level for both tests. The p-values for the Wald and Likelihood ratio tests, respectively, are 4.88e-05 and 4.4e-05. They are not the same because the tests assume different distributions.

1.1.3 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.

```
##
## Call:
## glm(formula = HD ~ AGE + SEX + SBP + 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 ***
## AGE          0.05664    0.01450    3.91  9.4e-05 ***
## SEXMALE      0.98987    0.14505    6.82  8.8e-12 ***
## SBP          0.01696    0.00236    7.18  7.0e-13 ***
## 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?

From the summary results below, we can observe that exhaustive search does not guarantee that the p-values will all be less then 0.05 for all remaining variables, which implies that it results in a different model than what we did for backwards elimination, with more variables.

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

```
##
## Call:
## glm(formula = HD ~ AGE + SEX + SBP + CHOL + FRW + CIG, family = binomial,
##     data = hd_data.f)
##
## 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
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: HD
##      LR Chisq Df Pr(>Chisq)
## AGE      17.6  1    2.7e-05 ***
## SEX      34.7  1    3.9e-09 ***
## SBP      42.3  1    7.8e-11 ***
## CHOL      8.9  1     0.0029 **
## FRW       2.3  1     0.1336
## CIG       4.0  1     0.0449 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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".

From the model in (ii), we can observe that all the coefficients are positive, meaning any unit increase in either age, SBP, CHOL, FRW or CIG will increase the likelihood of getting heart disease when controlling for all other variables. Among these numerical variables, it seems that AGE and SBP are the most important factors, meaning any variation has the most impact over increasing probability of heart disease.

It is also interesting to note the impact of the categorical variable SEX. Keeping all other factors constant, males have significantly more chance than females to have heart disease (91%!).

iv. What is the probability that Liz will have heart disease, according to our final model?

The chance of Liz having heart disease according to our final model is 3.46%.

1.2 Classification analysis

1.2.1 ROC/FDR

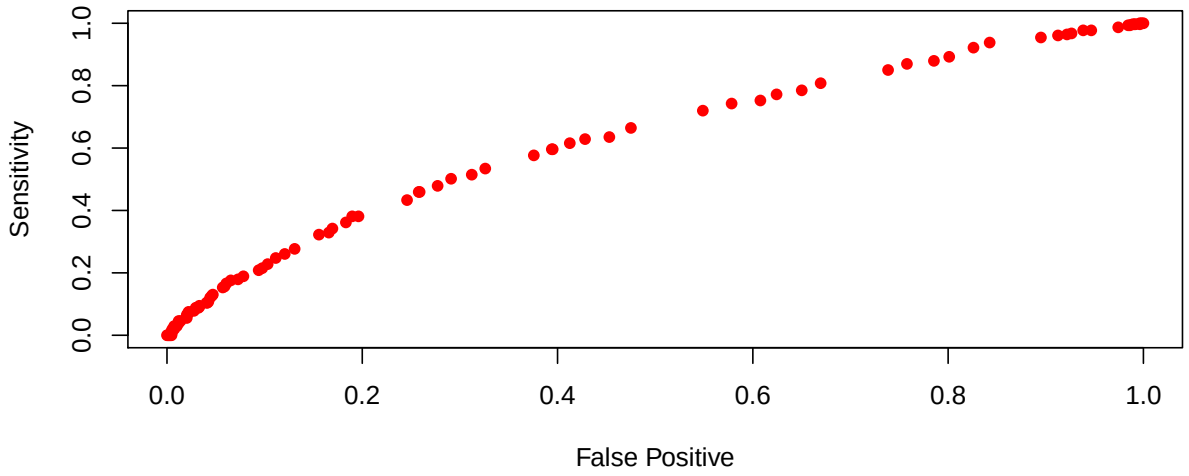
i. Display the ROC curve using `fit1`. Explain what ROC reports and how to use the graph. Specify the classifier such that the False Positive rate is less than .1 and the True Positive rate is as high as possible.

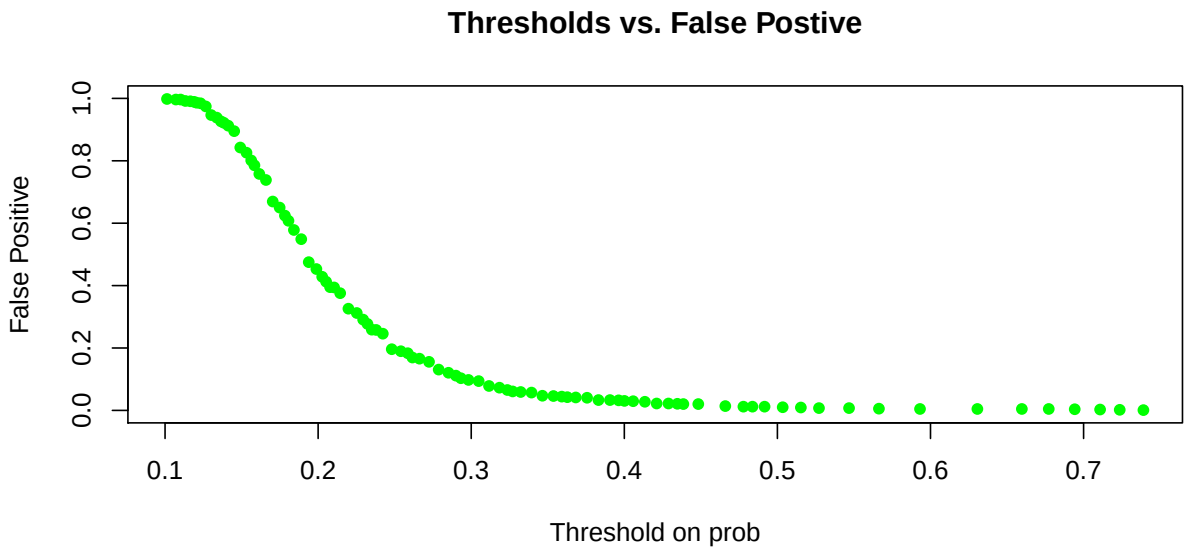
The ROC reports the pairs of sensitivity and specificity (or in the chart shown below, False Positives, which are 1-specificity). By showing the trade-off between increasing true positives and false positives in classification, it helps to establish the classifier once we define for the problem what is the maximum tolerance for False Negatives (or minimum for True Positives).

For a maximum False Positive of 0.1 and the highest possible Positive rate, we would need to choose a classifier of 0.3, according to the second graph below.

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

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



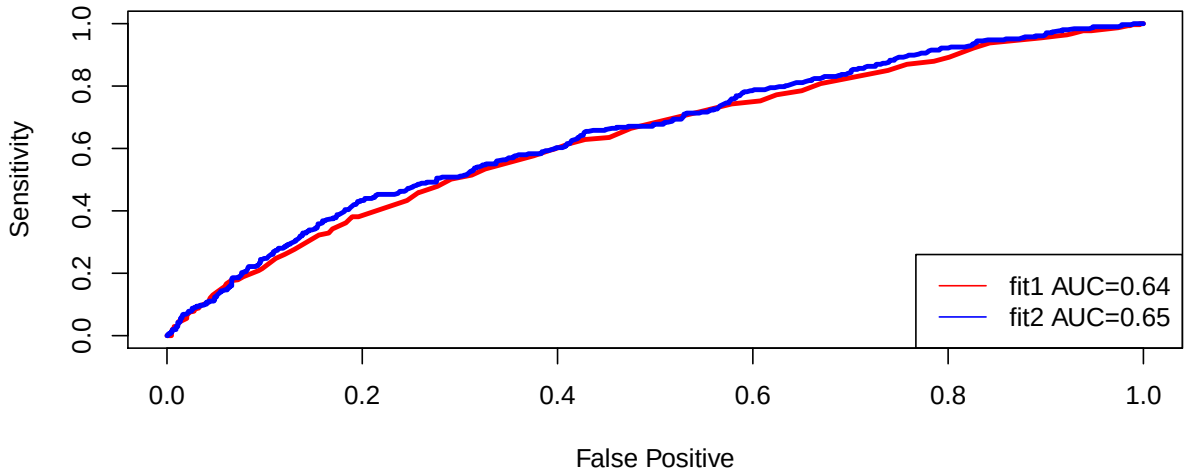


ii. Overlay two ROC curves: one from `fit1`, the other from `fit2`. Does one curve always contain the other curve? Is the AUC of one curve always larger than the AUC of the other one? Why or why not?

From the fitting of the two ROC curves, it seems that mostly the curve from fit2 contain that of fit1, which makes sense given fit1 is nested within fit2 and the larger number of variance in fit2 will generate less deviance (equivalent to residual square errors), and a slightly better accuracy.

```
## 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?

Below, you can find the confusion matrix, Positive Prediction value and Negative Prediction value for fit1 and fit2, respectively. If we were to Prioritize Positive Prediction values, fit2 would be more desirable, because it returns a higher proportion of real positives relative to the prediction of positives.

```
##
## fit1.pred    0    1
##         0 1075  298
##         1   11    9
```

```
## [1] 0.45
```

```
## [1] 0.783
```

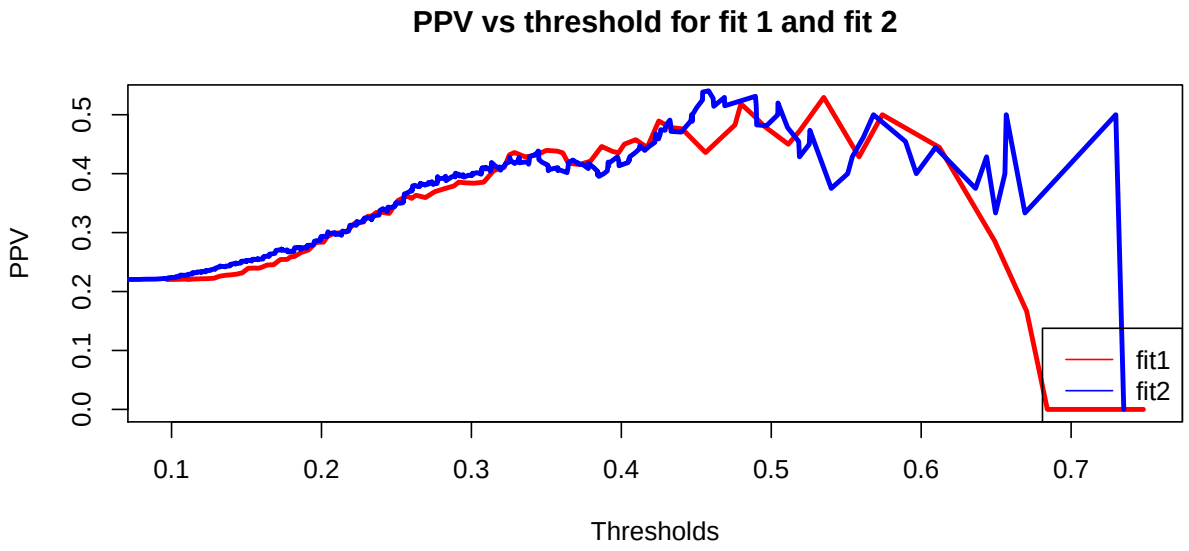
```
##
## fit2.pred    0    1
##         0 1073  294
##         1   13   13
```

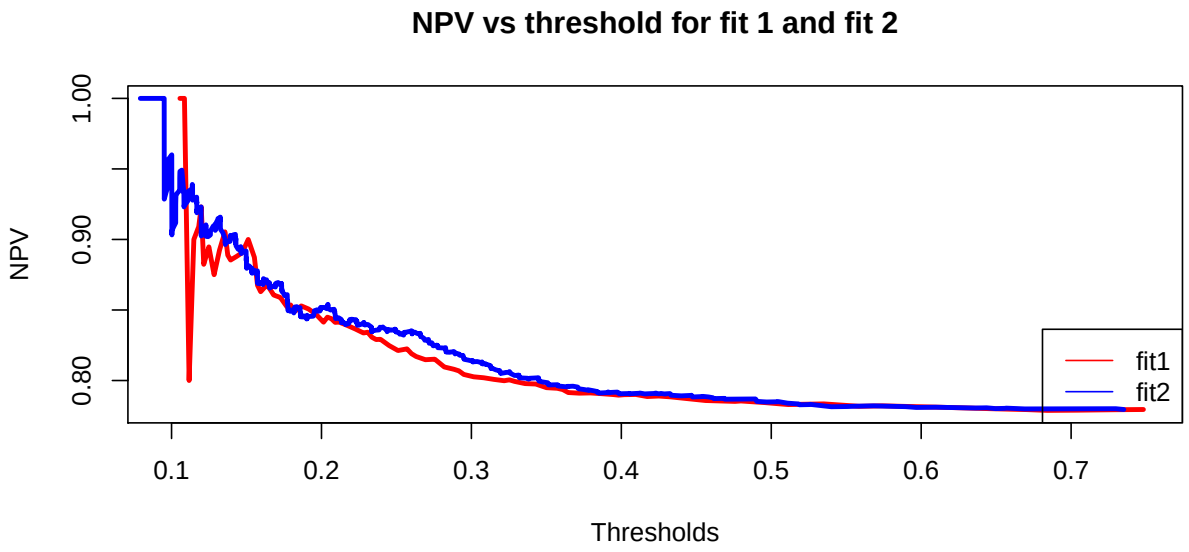
```
## [1] 0.5
```

```
## [1] 0.785
```

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.

Looking at the plots below and assuming that the set of positive and negative prediction values are the main concerns, i.e., the proportion of the actual positives that were classified as positives or as negatives, I would choose fit 2, because at any given threshold, the proportion of cases classified as positives and are true positives are mostly higher for fit 2, and the proportion of those classified as positives that were negative are mostly lower than fit 1 as well.





1.2.2 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 $a_{10}/a_{01} = 10$.

From the risk ratio: P(Y = 1 | x) > 0.1 / (1+0.1) = 0.09 logit > log (0.09/0.91) = -2.31

It follows that the linear boundary is: -2.31 <= -9.23 + 0.06*AGE* + *0.91*SEXMALE + 0.016*SBP* + *0.0045*FRW + 0.012*CIG

Or simplifying: 0 <= 6.62 + 0.06*AGE* + *0.91*SEXMALE + 0.016*SBP* + *0.0045*FRW + 0.012*CIG

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

```
## [1] 0.716
```

iii. How would you classify Liz under this classifier?

We saw before that Liz's odds of getting HD=1 was 0.0346. Because this is lower than 0.09, she would be classified as HD=0.

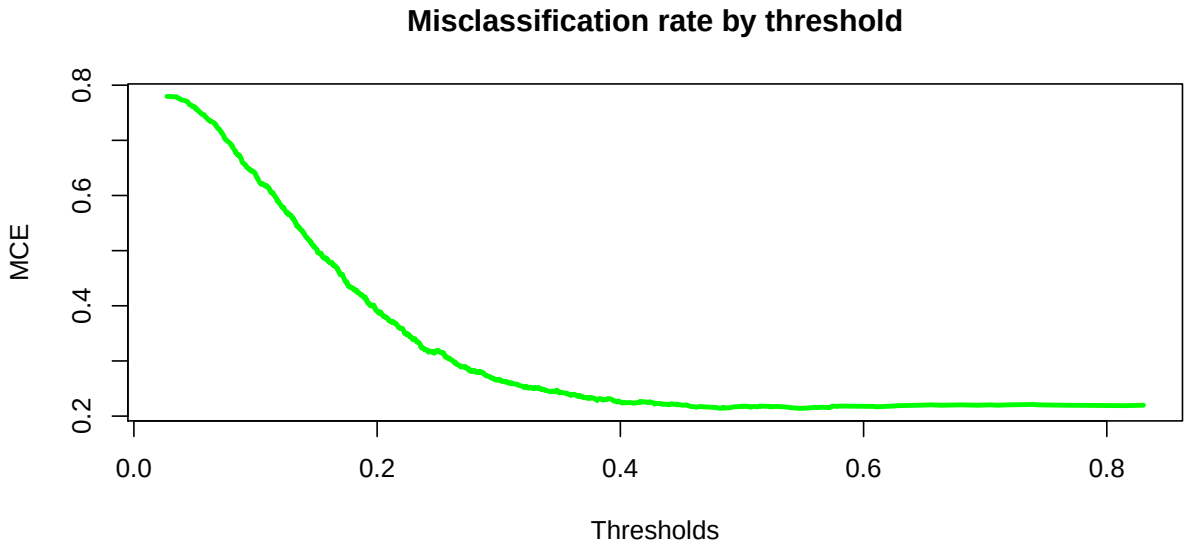
iv. Bayes rule gives us the best rule if we can estimate the probability of `HD-1` accurately. In practice we use logistic regression as our working model. How well does the Bayes rule work in practice? We hope to show in this example it works pretty well.

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

From the graph below, it can be observed that the Bayes rule works pretty well in practice, as we could judge from the elbow rule that ~0.1 would be a good cutoff to reduce misclassification error.

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

## Setting direction: controls < cases
```



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

We saw from question ii that using the risk ratio of 10 results in a weighted miscalssification error of 0.716.

vi. Use weighted misclassification error, and set $a_{10}/a_{01} = 1$. How well does the Bayes rule classifier perform?

By changing the risk ratio to 1, meaning the threshold will now become a cutoff at 50%, the weighted misclassification error actually reduces to 0.218, which is intuitive given that at 50% the classifier will treat every error equally, but if we care more about reducing negative prediction rates, then it makes more sense from the managerial standpoint to accept a higher misclassification error to compensate for a higher cost or risk ratio, of misclassifying a HD=1 to an HD=0 (i.e. not identify that the person has heart disease when he/she actually do).