Modern Data Mining, HW 4

Ziyi You Chenxi Leng Jihan Zhang

11:59 pm, 03/20, 2021

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 R Markdown / Knitr tips

You should think of this R Markdown file as generating a polished report, one that you would be happy to show other people (or your boss). There shouldn't be any extraneous output; all graphs and code run should clearly have a reason to be run. That means that any output in the final file should have explanations.

A few tips:

- Keep each chunk to only output one thing! In R, if you're not doing an assignment (with the `<-` operator), it's probably going to print something.

- If you don't want to print the R code you wrote (but want to run it, and want to show the results), use a chunk declaration like this: `{r, echo=F}`. Notice this is set as a global option.
- If you don't want to show the results of the R code or the original code, use a chunk declaration like: `{r, include=F}`
- If you don't want to show the results, but show the original code, use a chunk declaration like: `{r, results='hide'}`.
- If you don't want to run the R code at all use `{r, eval = F}`.
- We show a few examples of these options in the below example code.
- For more details about these R Markdown options, see the documentation.
- Delete the instructions and this R Markdown section, since they're not part of your overall report.

1.3 Review

Review the code and concepts covered in

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

1.4 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:

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

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

$$\begin{aligned}\mathcal{L}(\beta_0,\beta_1|\mathrm{D}ata) &= \mathbb{P}(\text{the outcome of the data})\\ &= \mathbb{P}((HD=1|SBP=152),(HD=0|SBP=110),(HD=0|SBP=154),(HD=1|SBP=160),(HD=0|S\\ &= \mathbb{P}(HD=1|SBP=152)\times\mathbb{P}(HD=0|SBP=110)\times\mathbb{P}(HD=0|SBP=154)\times\mathbb{P}(HD=1|SBP=160)\times\\ &= \frac{e^{\beta_0+152\beta_1}}{1+e^{\beta_0+152\beta_1}}\cdot\frac{1}{1+e^{\beta_0+110\beta_1}}\cdot\frac{1}{1+e^{\beta_0+154\beta_1}}\cdot\frac{e^{\beta_0+160\beta_1}}{1+e^{\beta_0+160\beta_1}}\cdot\frac{1}{1+e^{\beta_0+182\beta_1}}\end{aligned}$$

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 estimated logit function of `SBP` is logit $= -2.5456 + 0.014$SBP. Moreover, $\mathbb{P}(HD=1)$ is an increasing function of `SBP` since $\hat{\beta}_1 = 0.014 > 0$, which means that when `SBP` increases, the chance of being `HD` increases. The estimates that maximizes the likelihood function is called the Maximum Likelihood Estimators. In order to find them, we choose to maximize $\log(\mathcal{L}(\beta_\prime,\beta_\infty||\mathcal{D}))$, which is equivalent as minimizing the cross entropy.

iv. Evaluate the probability of Liz having heart disease.

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.

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.

	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?

The residual deviance of `fit2` is always smaller than that of `fit1` since we are adding more variables.

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?

As we can see from above results, the p-values for SEX from each test are different.

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.

Based on the above model with all variables, the variable with the largest p-value is `DBP`, so we will first kick out `DBP` and refit the model.

Now, the p-value for `FRW` is $0.1315 > 0.05$, so we will eliminate `FRW` and refit the model as follows.

Notice that the p-value of `CIG` is 0.0608, which is slightly higher than 0.05, but we still choose to elimiate it.

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?

First, we will find a model with smallest AIC through exhaustive search.

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

However, the exhaustive search does not guarantee that the p-values for all variables are less than .05. For example, `FRW` is not significant, so we will eliminate it.

Similarly, since the p-value for `CIG` is greater than .05, we will eliminate `CIG` for our final model as well.

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".

The final model we built in part(ii) shows that the probability of being HD increases 0.05664 as `AGE` increases by one unit with other factors fixed. When `SBP` increases one unit, the chance of HD increases 0.01696 while holding all other variables. Moreover, the probability of getting HD increases 0.00448 if the level of `CHOL` increases by one unit while fixing all other variables. Controlling all other factors, males have higher chance of getting HD compared to females.

A factor is important if the change of the variable value will change (increase/decrease) the probability of HD with all other factors in the model fixed.

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

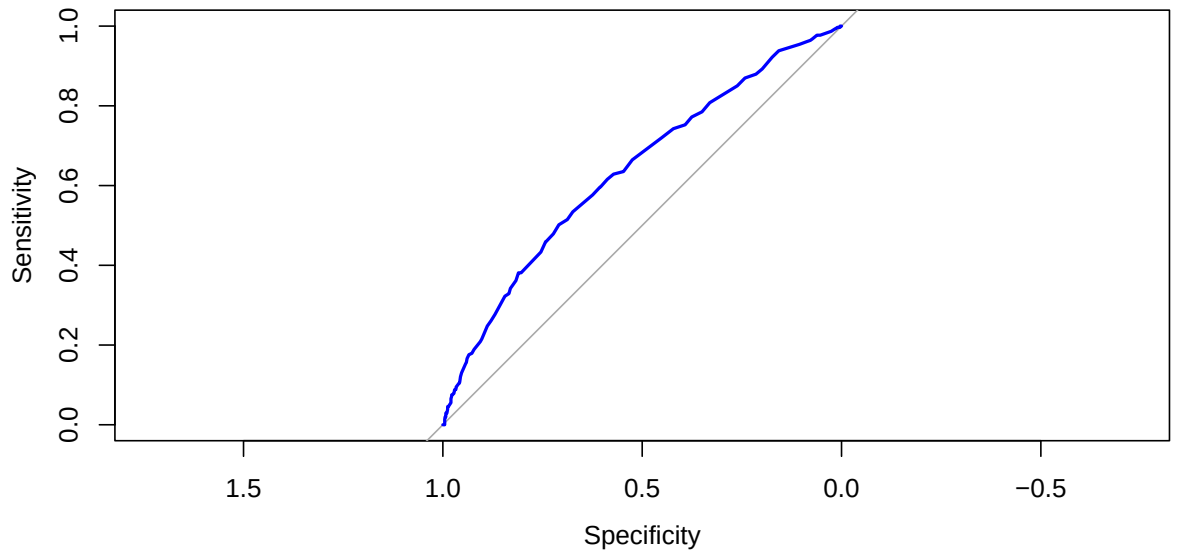
2.2 Classification analysis

2.2.1 ROC/FDR

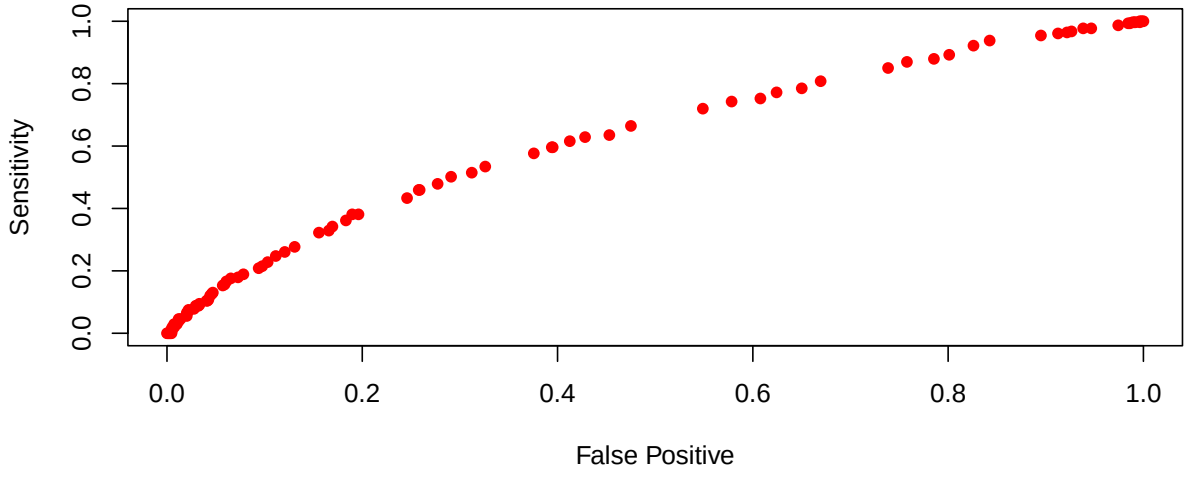
i. Display the ROC curve using `fit1`. Explain what ROC reports and how to use the graph. Specify the classifier such that the False Positive rate is less than .1 and the True Positive rate is as high as possible.

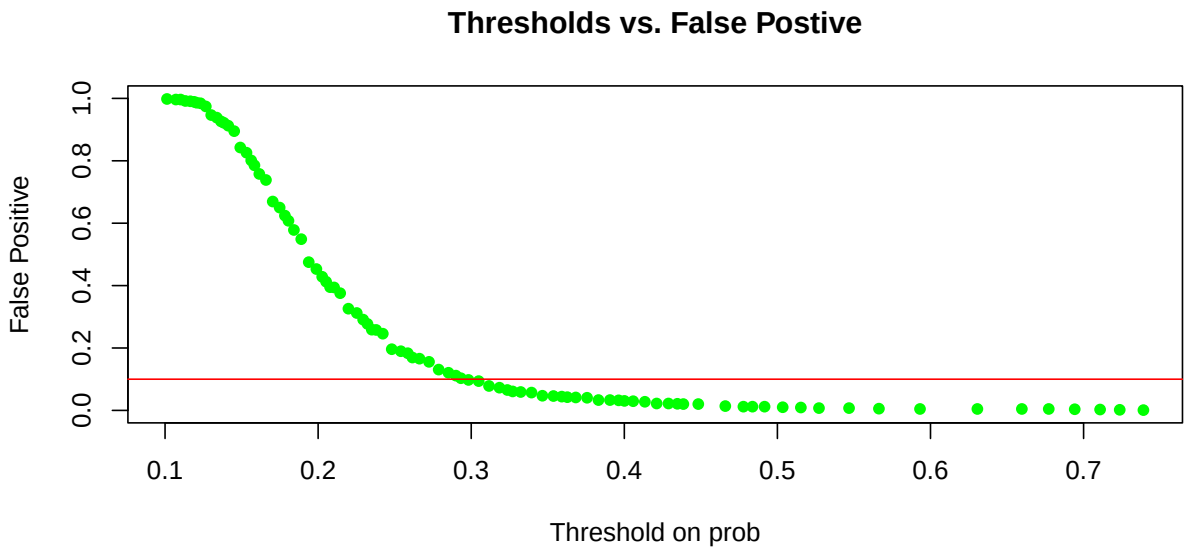
```
## Setting levels: control = 0, case = 1
```

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



The ROC curve shows that the higher the specificity, the lower the sensitivity. It can be used to choose classifiers. We want both the specificity and sensitivity to be as high as possible since we want both the the proportion of correct positive classification and the proportion of correct negative classification to be high. The balance between those two can be found using the ROC curve.





If we want our classifier to have False Positive rate less than .1 and the True Positive rate is as high as possible, we need to set the threshold to be larger than around 0.3.

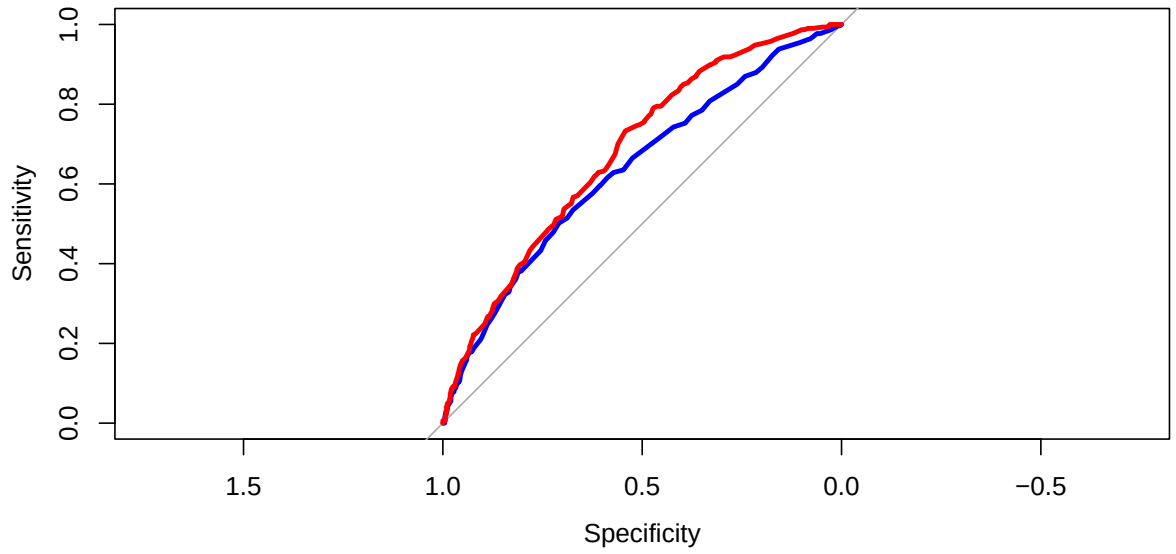
ii. Overlay two ROC curves: one from `fit1`, the other from `fit2`. Does one curve always contain the other curve? Is the AUC of one curve always larger than the AUC of the other one? Why or why not?

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

## Setting direction: controls < cases

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

## Setting direction: controls < cases
```

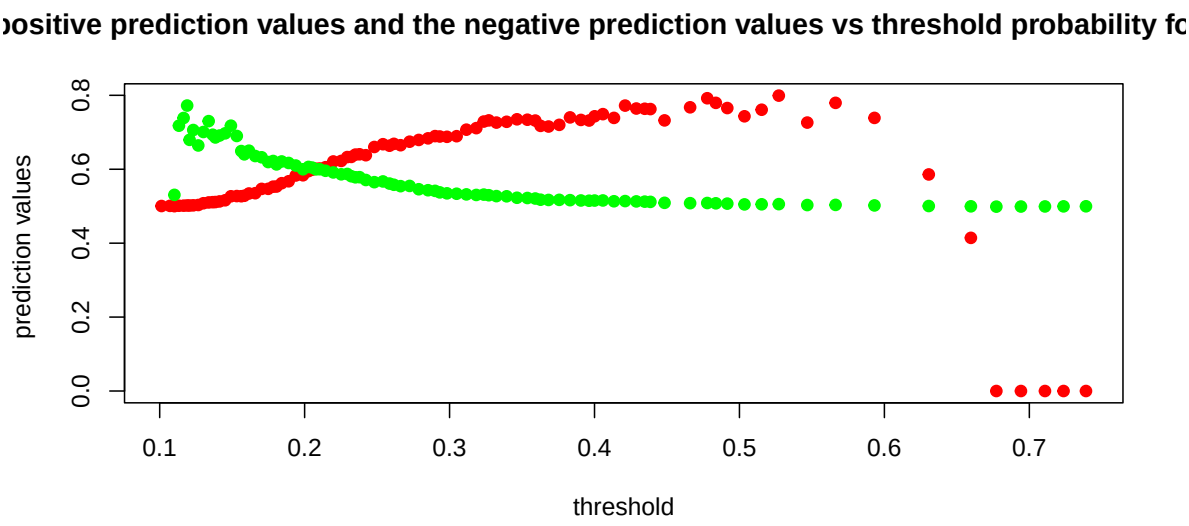


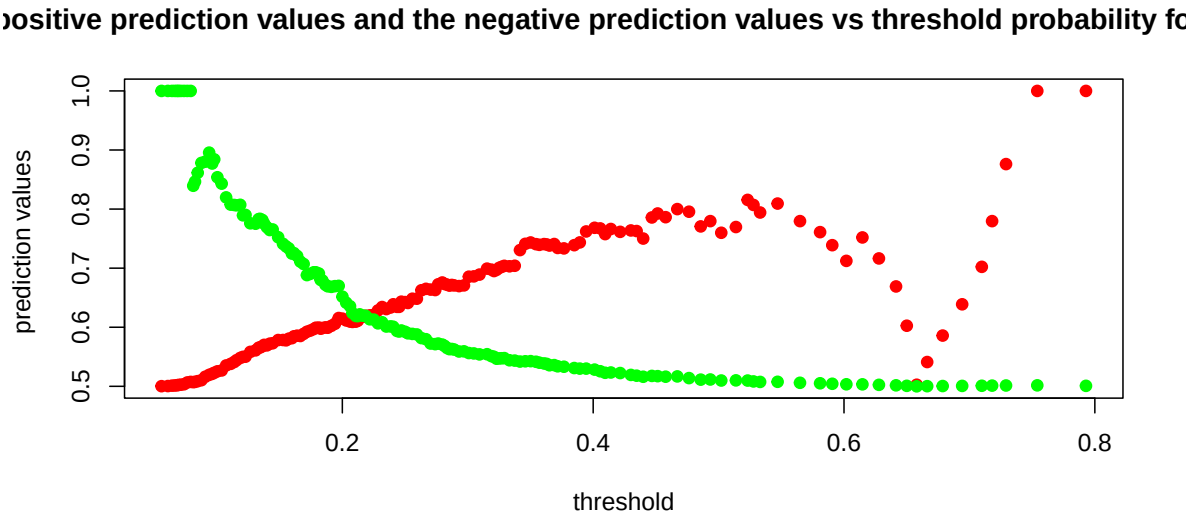
The blue line shows the ROC curve from `fit1` and the red line represents the ROC curve from `fit2`. Even though we have one more variable in `fit2`, the ROC curve from `fit2` does not always contain the ROC curve from `fit1`. However, the AUC of `fit2` is higher than that of `fit1`. With more variables, the model can have a better overall performance but we cannot tell if it will have higher proportion of correct positive classification, i.e. Sensitivity, with all fixed values of the proportion of correct negative classification, i.e. Specificity.

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?

As we can see from above two confusion matrices, the numbers of Positive Prediction Values for `fit1` and `fit2` are $\frac{9}{11+9} = 0.45$ and $\frac{17}{19+17} = 0.472$, respectively while those of Negative Prediction Values are $\frac{1075}{1075+11} = 0.783$ and $\frac{1067}{1067+19} = 0.786$. Therefore, if we prioritize the Positive Prediction values, we would like to use `fit2`.

iv. For `fit1`: overlay two curves, 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.





Note that the red dots are the positive prediction values and the green dots show the negative prediction values. I would like choose `fit2` so that the balance between the positive and negative prediction values can be achieved better.

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.

The final model we built from Part 1 is shown below

The logit for this final model can be written as

$$\begin{aligned}\text{logit}(P(HD = 1|x)) &= \log\left(\frac{P(HD = 1|x)}{P(HD = 0|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \\ &= -8.40872 + 0.05664 \cdot \text{AGE} + 0.01696 \cdot \text{SBP} + 0.00448 \cdot \text{CHOL} + 0.98987 \cdot \text{SEX}\end{aligned}$$

i. Write down the linear boundary for the Bayes classifier if the risk ratio of $a_{10}/a_{01} = 10$.

If the risk ratio of $a_{10}/a_{01} = 10$, the Bayes rule is thresholding over the $\hat{P}(Y = 1|x) > \frac{0.1}{(1+0.1)} = 0.0909$ and $\text{logit}(P(HD = 1|x)) > \log(\frac{0.0909}{1-0.0909}) = -2.3$

Then, the linear boundary for the Bayes classifier will be

$$\begin{aligned}\text{AGE} &> 107.85 - 0.299 \cdot \text{SBP} - 0.079 \cdot \text{CHOL} - 17.48 \cdot \text{SEX} \\ \text{SBP} &> 360.18 - 3.339 \cdot \text{AGE} - 0.264 \cdot \text{CHOL} - 58.37 \cdot \text{SEX} \\ \text{CHOL} &> 1363.55 - 12.643 \cdot \text{AGE} - 3.786 \cdot \text{SBP} - 220.95 \cdot \text{SEX} \\ \text{SEX} &> 6.171 - 0.057 \cdot \text{AGE} - 0.017 \cdot \text{SBP} - 0.0048 \cdot \text{CHOL}\end{aligned}$$

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

iii. How would you classify Liz under this classifier?

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.

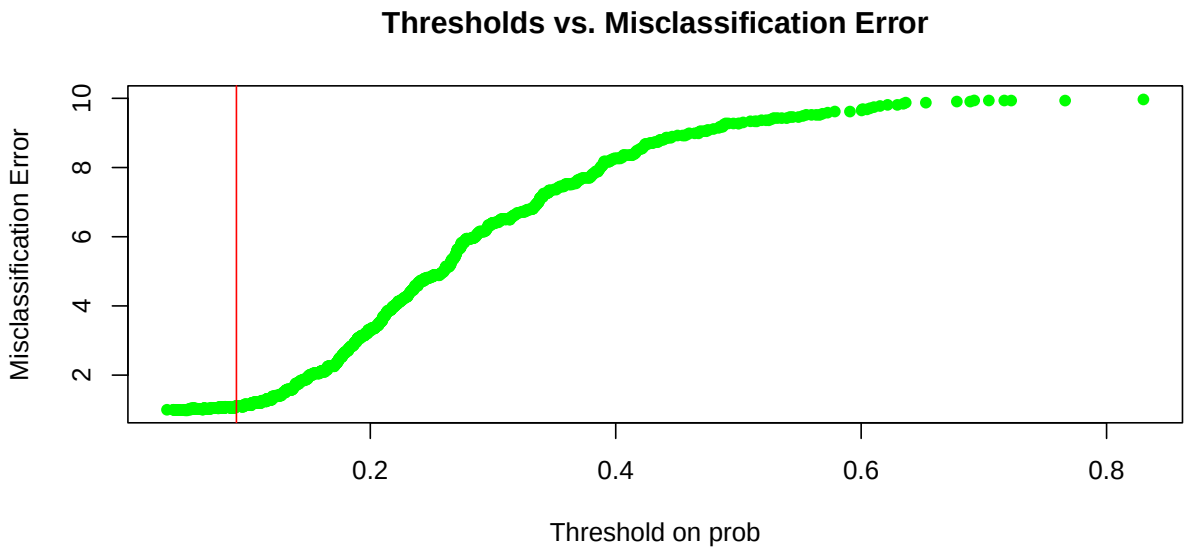
In practice, the Bayes rule can be used to determine the posterior conditional probability. In addtion to medical applications such as disease diagnosis, the Bayes rule can be used in finance in calculating or updating risk evaluation. As we have implemented in this example, if heart disease is related to age, then, applying Bayes' theorem, a person's age can be used to more accurately assess the probability that they have the heart disease, compared to the assessment of the probability of heart disease made without knowledge of the person's age.

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?

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

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

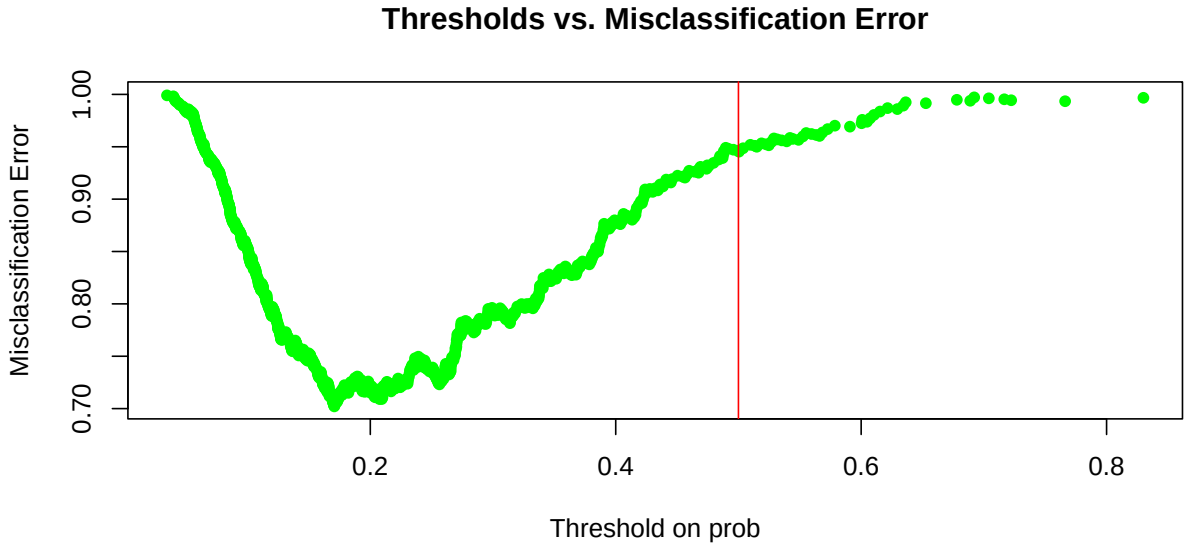


As we can see that the Bayes rule finds the optimal threshold that gives the lowest weighted misclassification error.

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

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

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



However, the Bayes rule does not work very well here to pick the smallest weighted misclassification error.

3 Part II: Project

3.1 Project Option 1 Credit Risk via LendingClub

3.2 Project Opetion 2 Diabetes and Health Management