Modern Data Mining, HW 4: Group 25

Brandon Kleinman, Philip Situmorang, Ben Sra Chongbanyatcharoen

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.

```
##     HD SBP
## 792  0 144
## 884  1 150
## 183  1 128
## 767  0 136
## 537  0 184
```

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

$$\begin{aligned}\mathcal{L}(\beta_0, \beta_1|\mathrm{D}ata) &= Prob(\text{the outcome of the data})\\ &= Prob((HD = 0|SBP = 144), (HD = 1|SBP = 150), (HD = 1|SBP = 128),\\ &(HD = 0|SBP = 136), (HD = 0|SBP = 184), \ldots)\\ &= \frac{1}{1 + e^{\beta_0 + 144\beta_1}} \cdot \frac{e^{\beta_0 + 150\beta_1}}{1 + e^{\beta_0 + 150\beta_1}} \cdot \frac{e^{\beta_0 + 128\beta_1}}{1 + e^{\beta_0 + 128\beta_1}} \cdot \frac{1}{1 + e^{\beta_0 + 136\beta_1}} \cdot \frac{1}{1 + e^{\beta_0 + 184\beta_1}} \cdots\end{aligned}$$

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.

Explanation on how MLE was obtained based on ii.

MLE are obtained by maximizing the likelihood function above. The intuitive explanation is that, the probability of getting the outcome of the data, which has already happened, should be as close as possible. Since the probability of an event occurring is modeled as a function with a value between 0 and 1, the best estimator (Betas) will be the ones that maximizes the value of the likelihood function.

Estimated logit function

The estimated logit function is: -3.65 + 0.0158 x SBP

Reference: output from glm()

```
##
## Call:
## glm(formula = HD ~ SBP, family = binomial(logit), 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
```

iv. Evaluate the probability of Liz having heart disease.

Now to estimate $P(HD = 1)$ for Liz, we plug in her `SBP=100` into the logistic regression from iii.

$$\hat{P}(HD = 1|SBP = 100) = \frac{e^{-3.65+0.01581\times SBP}}{1+e^{-3.65+0.01581\times ppSBP}} = \frac{e^{-3.65+0.01581\times 100}}{1+e^{-3.65+0.01581\times 100}} \approx 0.112$$

The probability that Liz has a heart disease is 0.112

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

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

From the perspective of getting the most accurate predictions, the categorical variable `SEX'` is most important to our model. This is because the absolute value of the estimator is largest, while it is also statistically significant.

We will create `fit2` with `SBP` and `SEX` as explanatory variables.

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?

In most cases, the residual deviance of `fit2` is likely to be smaller than that of `fit1`. This is because the residual deviance will if a variable added to the model have some predictive power. In the worst case, the residual deviance will stay the same as a new variable is added, if the newly added variable does not add any predictive power to the model. Such cases are rare.

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?

`AGE`, the added variable, is significant at 0.01 level in both Wald and Likelihood ratio tests. The p-values from each test are different. They are as the following:

- Wald test: `1.0e-10`
- Likelihood ratio test (Chi-squared): `3.8e-11`

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.

The model we created with the backward selection method is:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.4087	0.9086	-9.25	0.0000
SBP	0.0170	0.0024	7.18	0.0000
SEXMALE	0.9899	0.1451	6.82	0.0000
CHOL	0.0045	0.0015	3.00	0.0027
AGE	0.0566	0.0145	3.91	0.0001

This is the process that we went through before arriving at the model shown above:

1. Fit all variables
2. Eliminate DBP (which is not statistically significant and has the largest p-value)
3. Eliminate FRW (which is not statistically significant and has the largest p-value)
4. Eliminate CIG (which is not statistically significant and has the largest p-value)

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 does not guarantee that the p-values for all the remaining variables are less than .05 This is because this method only considers prediction accuracy as represented by AIC.

This is the model we arrived at by using AIC as the criterion for model selection:

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.3166	0.9714	-8.56	0.0000
AGE	0.0615	0.0148	4.16	0.0000
SEXFEMALE	-0.9113	0.1571	-5.80	0.0000
SBP	0.0160	0.0025	6.42	0.0000
CHOL	0.0045	0.0015	2.99	0.0028
FRW	0.0060	0.0040	1.51	0.1315
CIG	0.0123	0.0061	2.02	0.0437

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”.

Brief summary on important factors

The important factors relating to heart disease and their relationship with the *log likelihood function* are:

- **Age** - controlling for other variables, loglik is expected to increase by 0.06 on average, if age increases by 1
- **SEX (Gender)** - controlling for other variables, loglik is expected to decrease by 0.9 on average, if the person is female
- **SBP (Systolic blood pressure)** - controlling for other variables, loglik is expected to increase by 0.016 on average, if SBP increases by 1
- **CHO (Cholesterol level)** - controlling for other variables, loglik is expected to increase by 0.004 on average, if CHO increases by 1
- **FRW (age and gender adjusted weight)** - controlling for other variables, loglik is expected to increase by 0.006 on average, if FRW increases by 1
- **CIG (Self-reported number of cigarettes smoked per week)** - controlling for other variables, loglik is expected to increase by 0.012 on average, if CIG increases by 1

In other words, all variables apart from sex are positively correlated with the probability of getting heart disease. And males have a higher probability of getting heart disease than females on average, controlling for other variables.

Definition of "important factors"

Important factors are defined as variables that, when included into the model, create a model with the best prediction accuracy. (This is because our final model is selected using AIC as the sole criterion.)

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

Our final model predicts that the probability that Liz will have heart disease is 0.0346

2.2 Classification analysis

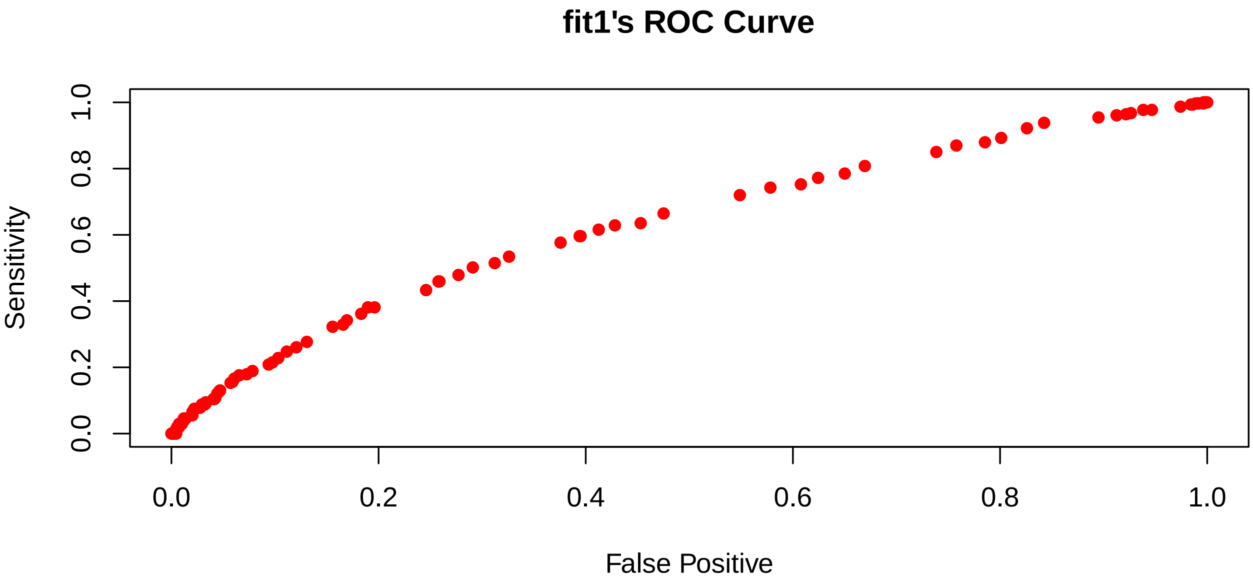
2.2.1 ROC/FDR

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.

Plot ROC Curve This is the ROC Curve of the model `fit1`. It shows the different possible combinations of true positive and false positive values that can be obtained using `fit1` under different threshold levels, from 0 to 1.

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

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



Specify classifier that maximizes TPR when FPR < 0.1 0.298 is the decision threshold for fit1 that keeps the false positive rate < 0.1 and maximizes the true positive rate subject to this constraint on the FPR. At the decision threshold of 0.298, the true positive rate is 0.215

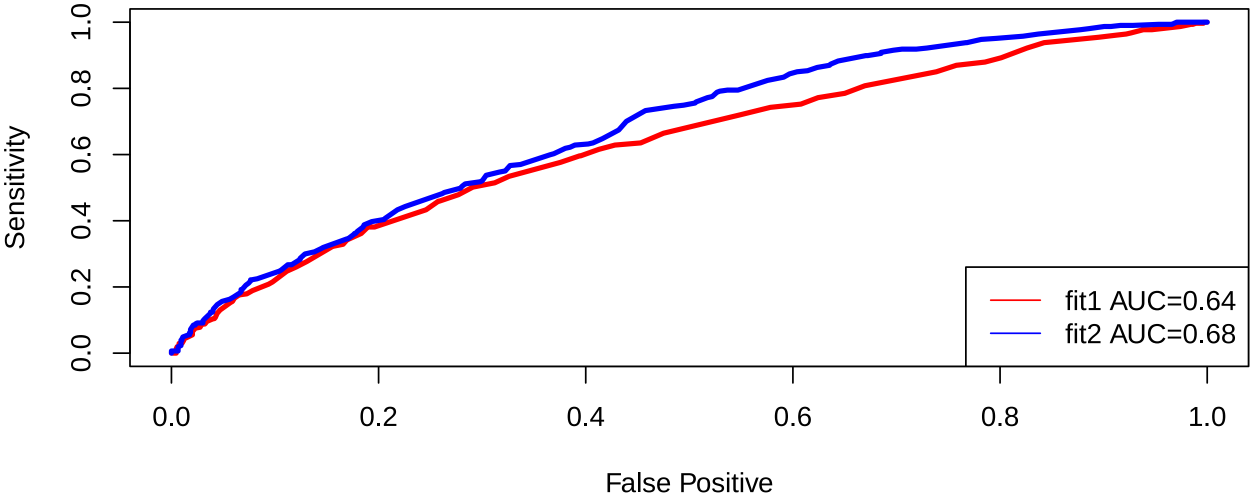
ii. Overlay two ROC curves: one from `fit1`, the other from `fit2`. Does one curve always contain the other curve? Is the AUC of one curve always larger than the AUC of the other one? Why or why not?

We can see that `fit1`'s ROC curve is inside `fit2`'s. This means that `fit2` is better in terms of overall performance, and that the AUC of `fit2`'s ROC curve is always higher than that of `fit1`'s.

ROC curves of fit1 and fit2

```
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?

The positive and negative prediction values for both models are as shown below.

`fit1`:

- Positive Prediction Value: 0.45
- Negative Prediction Value: 0.783

`fit2`:

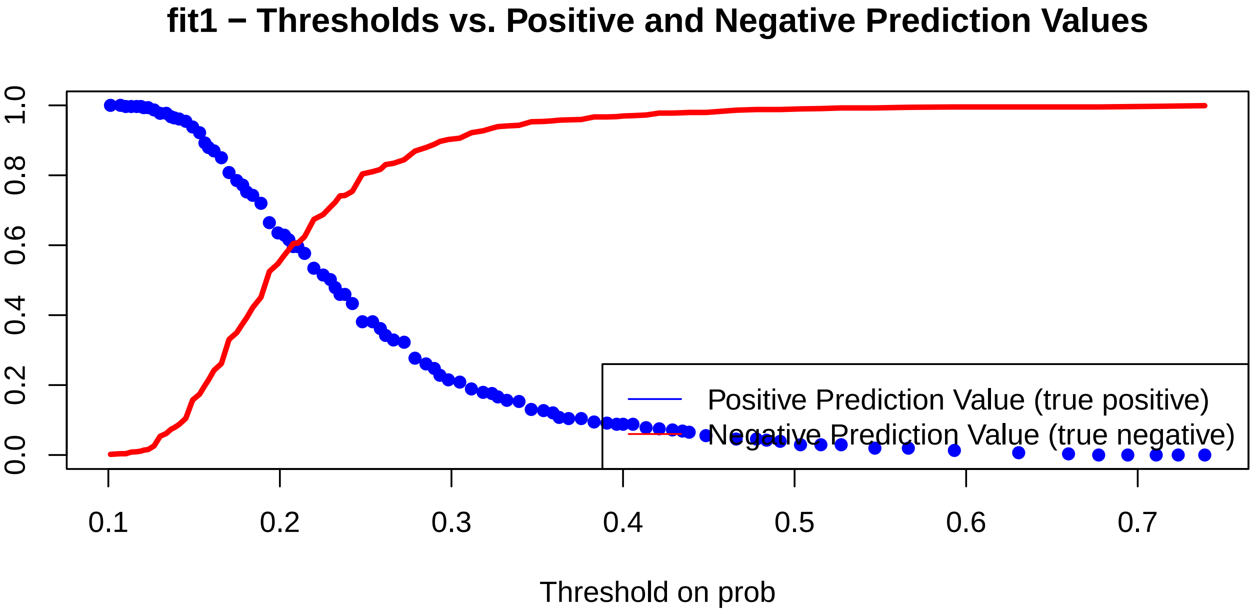
- Positive Prediction Value (true positive): 0.472
- Negative Prediction Value (true negative): 0.786

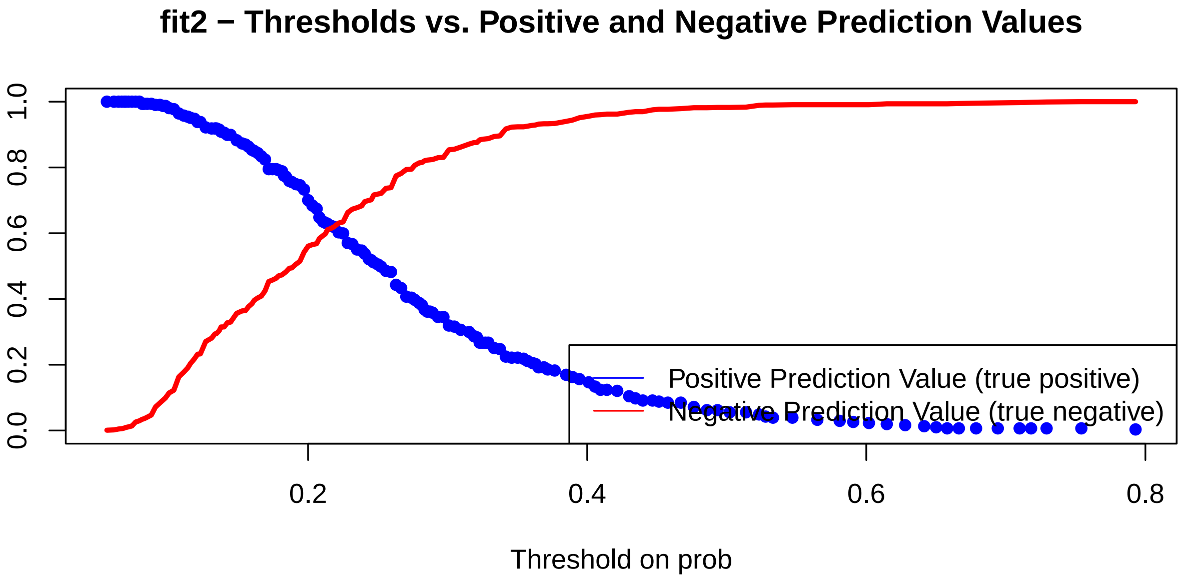
If we prioritize the Positive Prediction value (true positive), then `fit2` will be the more desirable model.

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.

Answer If the set of positive and negative prediction values are the concerns, we will choose `fit2`. This is because the sum of positive and negative prediction values for `fit2` is higher than that of `fit1` for every threshold.

Plots for both `fit1` and `fit2`





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

The linear boundary is:

$\widehat{HD} = 1$ if $.06153AGE - 0.91127SEXFEMALE + .01597SBP + .00449CHOL + .00604FRW + .01228CIG > 6.02$

Reference: calculation process

$$\hat{P}(Y = 1|x) > \frac{0.1}{(1 + 0.1)} = 0.0909$$

$$logit > \log(\frac{0.0909}{0.9090}) = -2.3$$

Recall that `logit` is:

$logit = -8.31658 + .06153AGE - 0.91127SEXFEMALE + .01597SBP + .00449CHOL + .00604FRW + .01228CIG$

Therefore,

$-8.31658 + .06153AGE - 0.91127SEXFEMALE + .01597SBP + .00449CHOL + .00604FRW + .01228CIG > -2.3$

$\widehat{HD} = 1$ if $.06153AGE - 0.91127SEXFEMALE + .01597SBP + .00449CHOL + .00604FRW + .01228CIG > 6.02$

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

** The weighted misclassification error is 0.714‘.

iii. How would you classify Liz under this classifier?

Under this classifier, Liz is still classified as "0" or negative for having heart disease.

This is because the threshold level is 0.0909, while Liz's probabiliy of having a heart disease, as predicted by our final model, is 0.0346, a level below that threshold.

Reference Recall that:

$$\hat{P}(Y = 1|x) > \frac{0.1}{(1 + 0.1)} = 0.0909$$

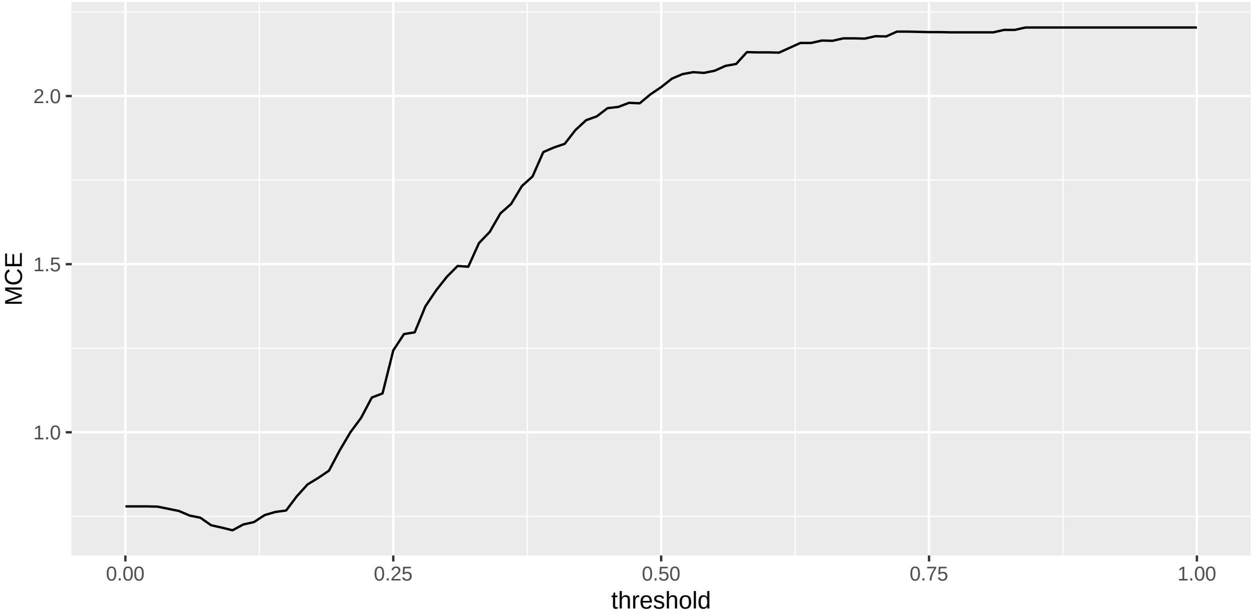
iv. Bayes rule gives us the best rule if we can estimate the probability of `HD-1` accurately. In practice we use logistic regression as our working model. How well does the Bayes rule work in practice? We hope to show in this example it works pretty well.

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 works well. With $a_{10}/a_{01} = 10$ we see that MCE is close to smallest at p = 0.0909, as expected.

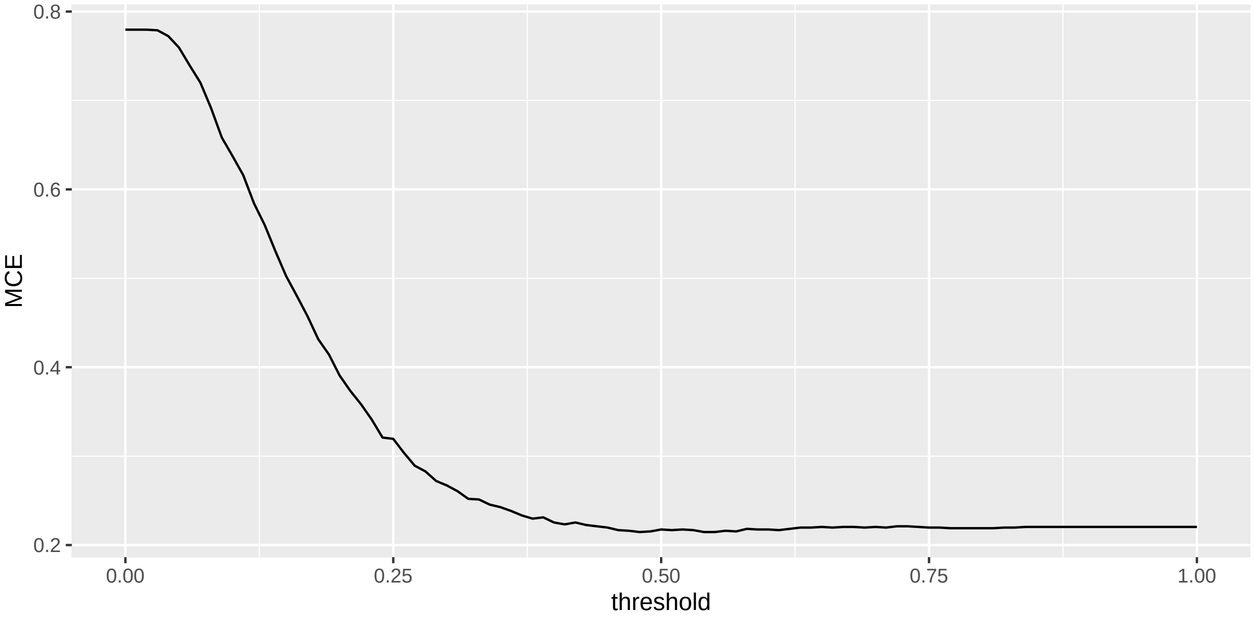
Reference: Curve between threshold level and MCE



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

The bayes rule works well. With $a_{10}/a_{01} = 1$ we see that MCE is close to smallest at p = 0.5, as expected.

Reference: Curve between threshold level and MCE



3 Part II: Project

3.1 Project Option 1 Credit Risk via LendingClub

3.2 Project Opetion 2 Diabetes and Health Management