Modern Data Mining, HW 4

Group Member Jenea Adams Group Member Annan Timon Group Member 3

11:59 pm, 03/19, 2023

Contents

1 Packages

2 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.

2.1 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`

2.2 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 in a chunk 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.

2.3 Review

Review the code and concepts covered in

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

2.4 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.

3 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.

```
   0    1 
1095  311 
```

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

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
```

3.1 Identify risk factors

3.1.1 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.

[sample]:[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.

$$P(HD=1|SBP) = \frac{e^{\beta_0+\beta_1 SBP}}{1+e^{\beta_0+\beta_1 SBP}}$$

$$P(HD=0|SBP) = 1 - P(HD=1|SBP) = \frac{1}{1+e^{\beta_0+\beta_1 SBP}}$$

$$\begin{aligned}
\mathcal{L}(\beta_0,\beta_1|\mathrm{D}ata) &= Prob(\text{the outcome of the data}) \\
&= Prob((HD=0|SBP=118),(HD=0|SBP=140),(HD=0|SBP=160),(HD=1|SBP=165),(HD= \\
&= Prob(HD=0|SBP=118)\times Prob(HD=0|SBP=140)\times Prob(HD=0|SBP=160)\times Prob(HD=1|S \\
&= \frac{1}{1+e^{\beta_0+118\beta_1}}\cdot\frac{1}{1+e^{\beta_0+140\beta_1}}\cdot\frac{1}{1+e^{\beta_0+160\beta_1}}\cdot\frac{e^{\beta_0+165\beta_1}}{1+e^{\beta_0+165\beta_1}}\cdot\frac{e^{\beta_0+215\beta_1}}{1+e^{\beta_0+215\beta_1}}
\end{aligned}$$

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 MLE are obtained numerically by maximizing the log likelihood function are above to estimate the unknown $\beta' sgiventhedata$

iv. Evaluate the probability of Liz having heart disease.

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.*

The probability of Liz having heart disease with an SBP of 110 is 12.8%

3.1.2 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`.

`SEX` is the single cariable that is the most important to add given the lowest AIC for that model

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)	-1.6260	0.0999	-16.28	0.0000
SEXMALE	0.6917	0.1320	5.24	0.0000

Call: glm(formula = HD ~ SEX, family = binomial, data = hd_data.f)

Deviance Residuals: Min 1Q Median 3Q Max
-0.814 -0.814 -0.599 -0.599 1.900

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.6260 0.0999 -16.28 < 2e-16 ***SEXMALE 0.6917 0.1320 5.24 1.6e-07*** — 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: 1441.2 on 1391 degrees of freedom AIC: 1445

Number of Fisher Scoring iterations: 4

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

The residual deviance of `fit2`, which is the smaller model, may not be consistently smaller than `fit1`. This is because the residual deviance is influenced by the the features in the proposed model compared to the saturated model, and while fit2 may exhibit the best feature to add to the model sequentially, it may not necessarily be the best fit for the data overall.

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?

Wald test

the pvalue for SEX is 3.8e-11

The confidence intervals do not contain 0 for any of the coefficients, so we know that the true betas for the variables are not 0.

Likelihood ratio test

the pvalue for SEX is 3.8e-11, which is the same as the type II test

the pvalue for this likelihood ratio test is 1.98e-22, which is still significant, but different as it's based of the entire model comparison of deviance residuals

3.1.3 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.

DBP has the highest p-value, so we will remove it.

FRW has the highest p-value, so we will remove it.

CIG has the highest p-value, so we will remove 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?

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

Exhaustive search dos not guarantee that the p-values for all the remaining variable are less than 0.05. The final model here is not the same as the model from backward selection

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".

AGE, SBP, CHOL, CIG are all positively related to the chance of a HD because of their positive Beta coefficients. This means that as these features increase, the chance of HD increases as well. Males have an increased risk of HD compared to females.

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

The final model predicts that the probability that Liz has heart disease is 4.9%

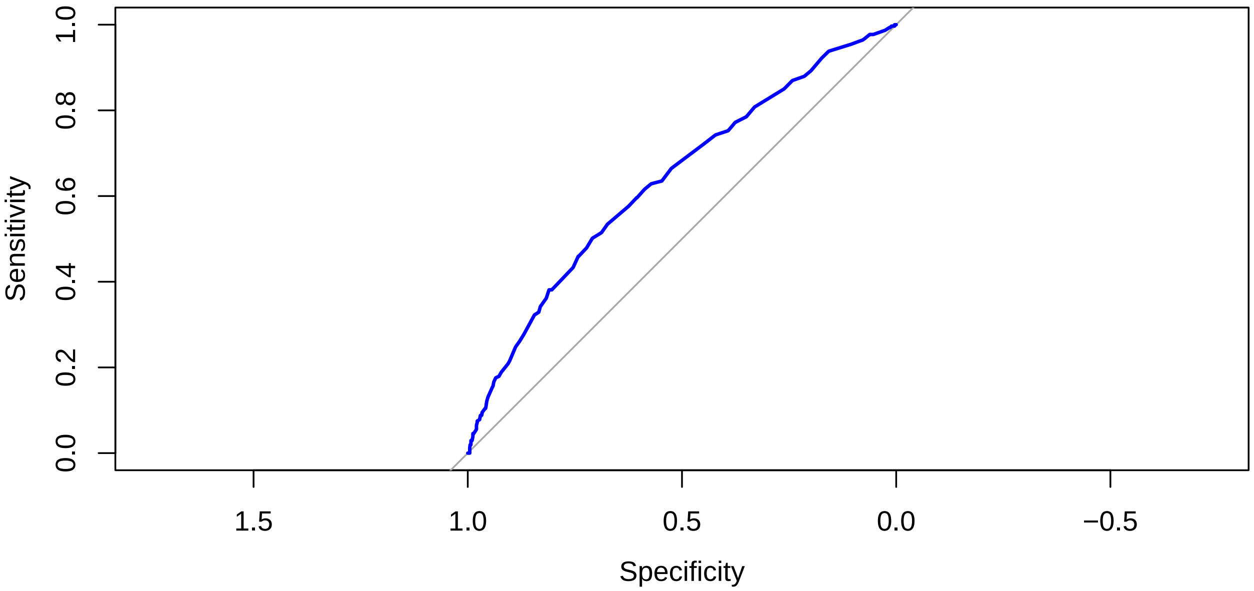
3.2 Classification analysis

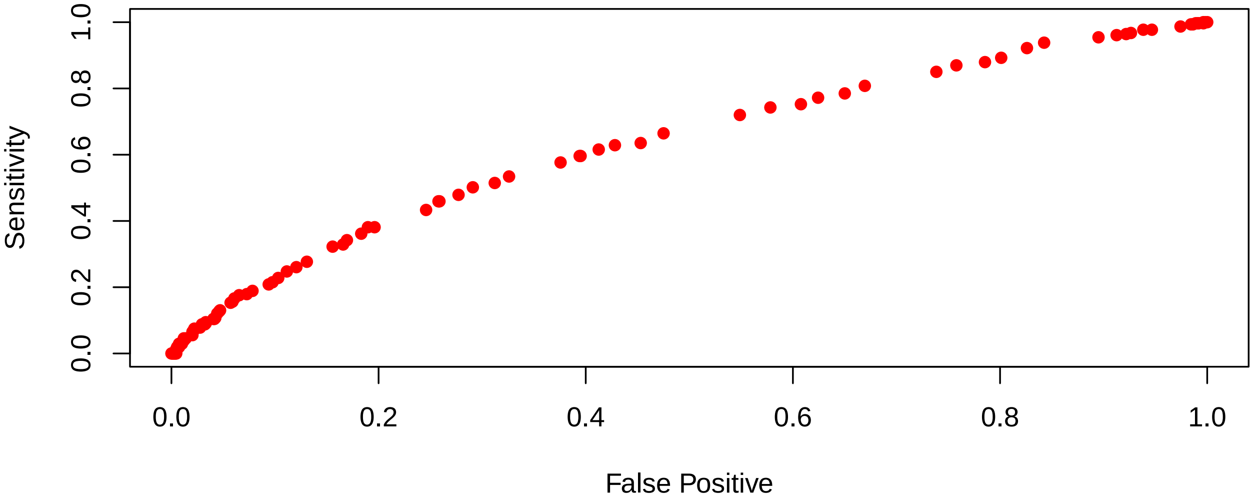
3.2.1 ROC/FDR

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
```

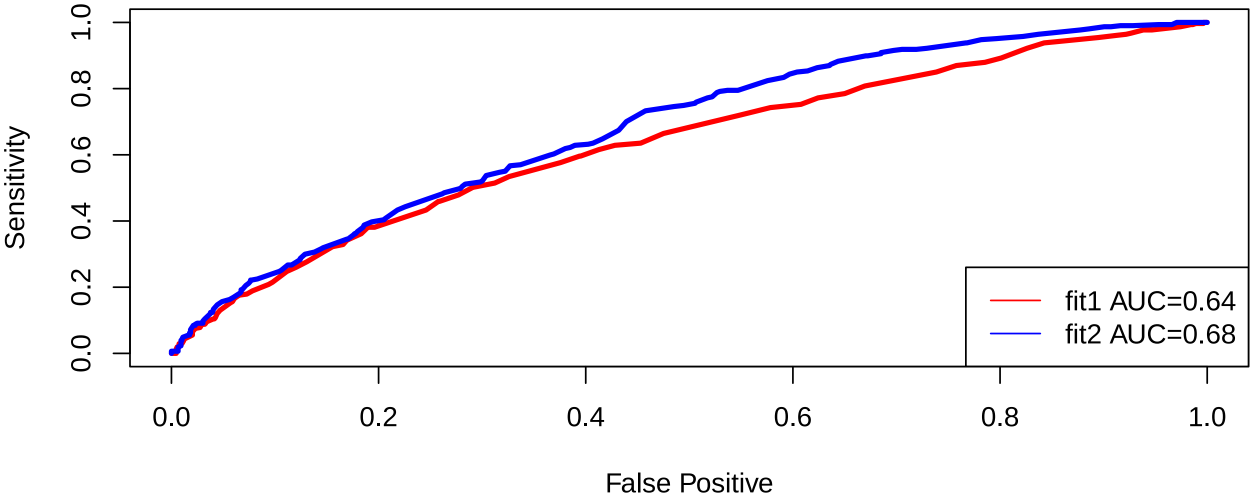




ii. Overlay two ROC curves: one from `fit1`, the other from `fit2`. Does one curve always contain the other curve? Is the AUC of one curve always larger than the AUC of the other one? Why or why not?

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

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



The AUC of `fit2` is generally larger than that of `fit1`, even with potential crossover in lower left of he plot.

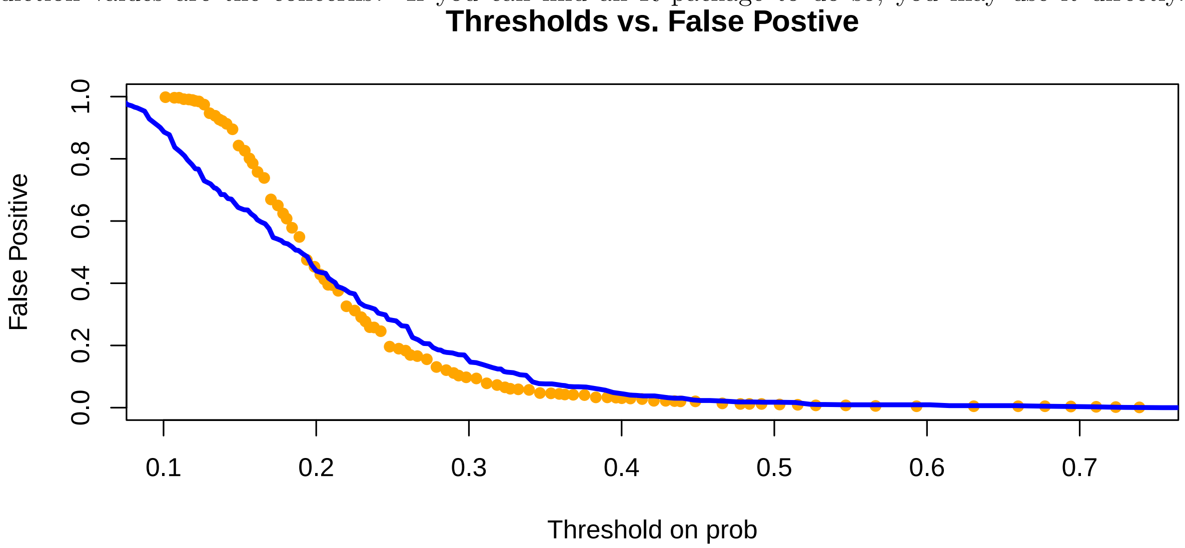
iii. Estimate the Positive Prediction Values and Negative Prediction Values for `fit1` and `fit2` using .5 as a threshold. Which model is more desirable if we prioritize the Positive Prediction values?

Here we see the Positive Prediction Values and Negative Prediction Values for `fit1` are 0.45 and 0.783, respectively. They are calculated manually and confirmed with the confusionMatrix function.

Here we see the Positive Prediction Values and Negative Prediction Values for `fit2` are 0.472 and 0.786, respectively. They are calculated manually and confirmed with the confusionMatrix function.

Fit2 is more desireable if we prioritze positive prediction because this is 0.472 compared to fit1's 0.45

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.



I would chose fit2 because it has a lower false positive rate than fit1 at lower thresholds.

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

If the cost of misclassifying "1" to "0" is 10 times of that of misclassifying "0" to "1", then the Bayes rule is thresholding over the

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

or

$$logit > \log(\frac{0.0909}{0.909}) = -2.3$$

The linear boundary is

$\hat{HD} = 1$ if $0 <$ 0.06153 AGE +0.91127 SEX-MALE +0.01597 SBP +0.00449 CHOL +0.006043 FRW +0.01228 CIG -6.93

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

The weighted misclassification error for this given risk ratio is 0.714

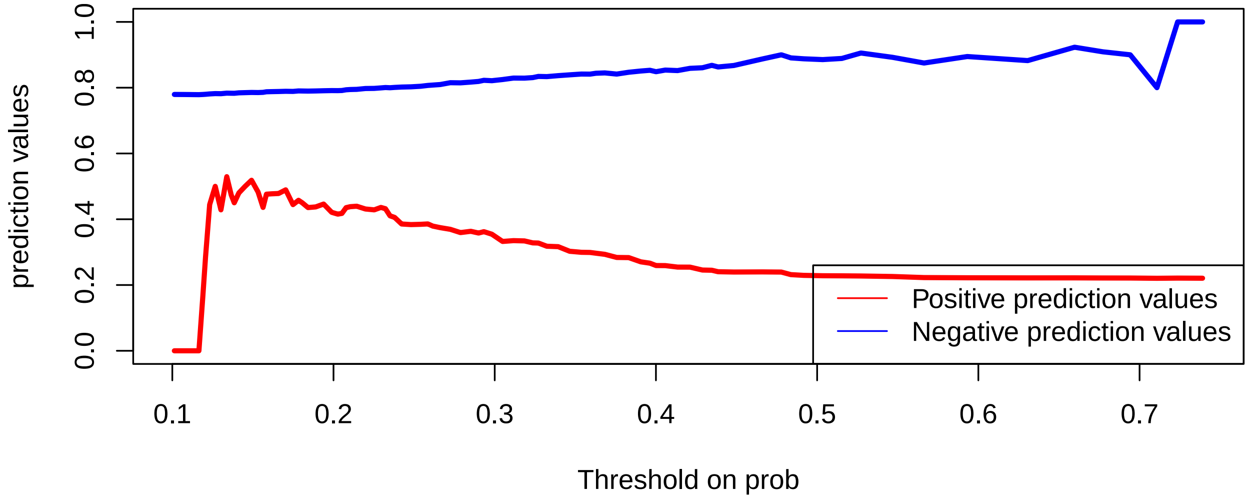
iii. How would you classify Liz under this classifier?

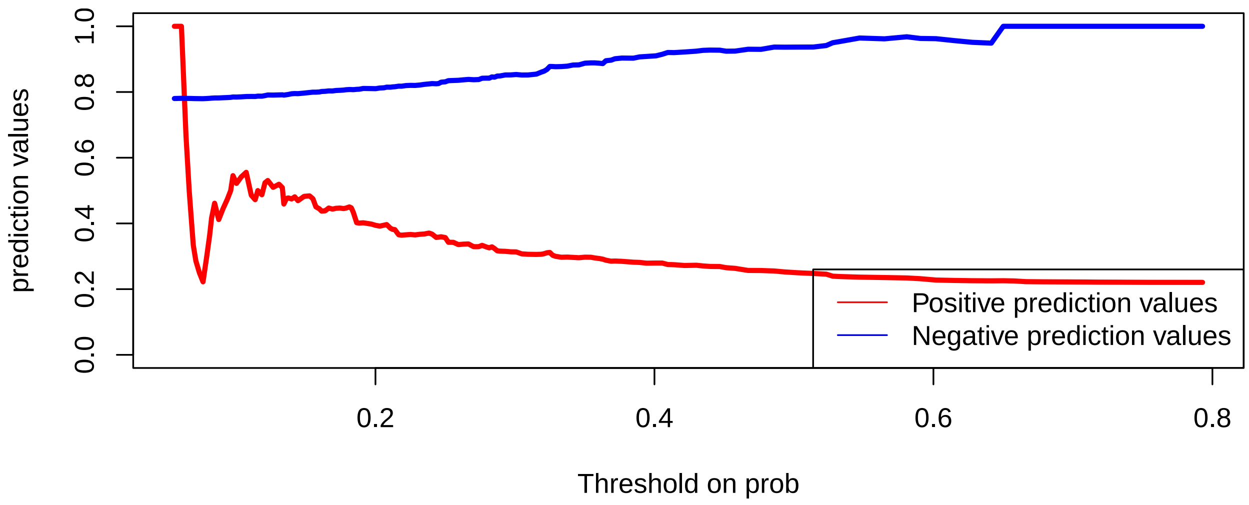
$$0.019 \times (110) + 0.903 \times (FEMALE) > log(10) + 4.57$$

2.09 is not greater than 5.57, so she is assigned a 0 meaning no heart disease.

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?

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

If the cost of misclassifying “1” to “0” is 1 times of that of misclassifying “0” to “1”,then the Bayes rule is thresholding over the

$$\hat{P}(Y = 1|x) > \frac{1}{(1+1)} = 0.5$$

or

$$logit > \log(\frac{05}{0.5}) = 0$$

The linear boundary is

$\hat{HD} = 1$ if $0 < 0.06153$ AGE $+0.91127$ SEX-MALE $+0.01597$ SBP $+0.00449$ CHOL $+0.006043$ FRW $+0.01228$ CIG -9.227

the weighted misclassification error is 2.05

4 Part II: Project

4.1 Project Option 1 Credit Risk via LendingClub

4.2 Project Opetion 2 Diabetes and Health Management