Modern Data Mining, HW 4

Annie Vo Sarah Hayward Jessica Brown

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.

```
   0    1
1095  311
```

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

```
# using the comment="     ", we get rid of the ## in the output.
summary(hd_data.f)
```

```
       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
```

```
row.names(hd_data.f) <- 1:1393
set.seed(1)
indx <- sample(1393, 5)
hd_data.f[indx, ]
```

```
     HD AGE    SEX SBP DBP CHOL FRW CIG
1017  0  52 FEMALE 152  90  207 108   0
679   0  46 FEMALE 116  70  237  93   0
129   0  46   MALE 154 100  145 107   0
930   0  53 FEMALE 125  85  255  74   0
471   0  55   MALE 136  96  250 102   0
```

```
set.seed(1)
hd_data.f[sample(1393, 5), ]
```

```
     HD AGE    SEX SBP DBP CHOL FRW CIG
1017  0  52 FEMALE 152  90  207 108   0
679   0  46 FEMALE 116  70  237  93   0
129   0  46   MALE 154 100  145 107   0
930   0  53 FEMALE 125  85  255  74   0
471   0  55   MALE 136  96  250 102   0
```

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
## 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.

$P(HD = 1 \mid SBP = 152) \times P(HD = 0 \mid SBP = 110) \times P(HD = 0 \mid SBP = 154) \times P(HD = 1 \mid SBP = 160) \times P(HD = 0 \mid SBP = 182) = \frac{e^{\beta_0+152\beta_1}}{1+e^{\beta_0+152\beta_1}} \cdot \frac{1}{1+e^{\beta_0+110\beta_1}} \cdot \frac{1}{1+e^{\beta_0+154\beta_1}} \cdot \frac{e^{\beta_0+160\beta_1}}{1+e^{\beta_0+160\beta_1}} \cdot \frac{1}{1+e^{\beta_0+182\beta_1}}$

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.

```
fit1 <- glm(HD~SBP, hd_data.f, family=binomial(logit))
summary(fit1, results=TRUE)
```

```
##
## 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
```

$logit = -3.65489 + 0.01581SBP$ $P(HD = 1 \mid SBP) = \frac{e^{-3.65489+0.01581SBP}}{1+e^{-3.65489+0.01581SBP}}$ The MLE is obtained through maximizing the likelihood function in ii above.

iv. Evaluate the probability of Liz having heart disease. $P(HD = 1 \mid SBP = 110) = \frac{e^{-3.65489+0.01581(110)}}{1+e^{-3.65489+0.01581(110)}} \approx 0.128331772$

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

```
fit1 <- glm(HD~SBP, hd_data.f, family=binomial)
summary(fit1)
fit1.1 <- glm(HD~SBP + AGE, hd_data.f, family=binomial)
summary(fit1.1)
# you will need to finish by adding each other variable
# fit1.2...
fit1.2 <- glm(HD~SBP + SEX, hd_data.f, family=binomial)
```

```
summary(fit1.2)
fit1.3 <- glm(HD~SBP + DBP, hd_data.f, family=binomial)
summary(fit1.3)
fit1.4 <- glm(HD~SBP + CHOL, hd_data.f, family=binomial)
summary(fit1.4)
fit1.5 <- glm(HD~SBP + FRW, hd_data.f, family=binomial)
summary(fit1.5)
fit1.6 <- glm(HD~SBP + CIG, hd_data.f, family=binomial)
summary(fit1.6)
```

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

The SEX variable would be the most important to add.

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.

```
## How to control the summary(fit2) output to cut some junk?
## We could use packages: xtable or broom.
## SEX has the highest z value and smallest p value
library(xtable)
options(xtable.comment = FALSE)
fit2 <- glm(HD~SBP + SEX, hd_data.f, family=binomial)
xtable(fit2)
```

	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

```
summary(fit2)
```

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

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

Yes because we are adding in a variable that explains more of the deviance. However, some variables might not decrease the residual deviance by much because they are not good prediction 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?

```
summary(fit2)
```

```
##
## 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
```

```
confint.default(fit2)
```

```
##               2.5 %  97.5 %
## (Intercept) -5.3341 -3.8064
## SBP          0.0142  0.0233
## SEXMALE      0.6295  1.1773
```

```
Anova(fit2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: HD
##     LR Chisq Df Pr(>Chisq)
## SBP     67.5  1    < 2e-16 ***
## SEX     43.7  1    3.8e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: The added variable of SEX is significant at the 0.01 level for both tests, with different p values of 0.0000000001 for the Wald test and .000000000038 for the Likelihood ratio test.

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.

```
fit3 <- glm(HD~SBP + AGE+ SEX + DBP + CHOL + FRW + CIG, hd_data.f, family=binomial)
summary(fit3)
#DBP has the highest p value of 0.70594 which is not significant at the 0.05 level
fit3.1 <- update(fit3, .~. -DBP)
summary(fit3.1)
#FRW has the highest p value of 0.1315 which is not significant at the 0.05 level
fit3.2 <- update(fit3.1, .~. -FRW)
summary(fit3.2)
#CIG has the highest p value of 0.0608 which is not significant at the 0.05 level
fit3.3 <- update(fit3.2, .~. -CIG)
summary(fit3.3)
#All variables (SBP, AGE, SEX, and CHOL) are now significant at the 0.05 level
```

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?

```
Xy_design <- model.matrix(HD ~.+0, hd_data.f)
Xy <- data.frame(Xy_design, hd_data.f$HD)

fit.all <- bestglm(Xy, family = binomial, method = "exhaustive", IC="AIC", nvmax = 10)
```

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

```
fit.all$BestModel
```

```
##
## Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Coefficients:
## (Intercept)          AGE      SEXMALE          SBP         CHOL          FRW
##    -9.22786      0.06153      0.91127      0.01597      0.00449      0.00604
##         CIG
##     0.01228
##
## Degrees of Freedom: 1392 Total (i.e. Null);  1386 Residual
## Null Deviance:       1470
## Residual Deviance: 1340  AIC: 1360
```

```
fit4 <-glm(HD~AGE+SEX+SBP+CHOL+FRW+CIG, family=binomial, data=hd_data.f)
summary(fit4)
```

```
##
## 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
```

```
#Dropping FRW because it is not significant at the 0.05 level
fit4.1 <- update(fit4, .~. -FRW)
summary(fit4.1)
```

```
##
## Call:
## glm(formula = HD ~ AGE + SEX + SBP + CHOL + CIG, family = binomial,
##     data = hd_data.f)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.754  -0.729  -0.554  -0.344   2.447
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.70228    0.92683   -9.39  < 2e-16 ***
## AGE          0.06136    0.01475    4.16  3.2e-05 ***
## SEXMALE      0.88575    0.15579    5.69  1.3e-08 ***
## SBP          0.01709    0.00237    7.20  5.9e-13 ***
## CHOL         0.00440    0.00150    2.93   0.0033 **
## CIG          0.01136    0.00606    1.87   0.0608 .
## ---
## 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: 1345.5  on 1387  degrees of freedom
## AIC: 1358
##
## Number of Fisher Scoring iterations: 4
```

```
#Dropping CIG because it is not significant at the 0.05 level
fit4.2 <- update(fit4.1, .~. -CIG)
summary(fit4.2)
```

```
##
## 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
```

The exhaustive search does not guarantee that the p values for all the remaining variables are less than 0.05. For example, the p value for FRW is 0.1315 which is greater than 0.05. We also dropped CIG after dropping FRW because the p value was also greater than 0.05 at 0.0608. The model here is the same as the one from backwards induction after dropping the insignificant variables.

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, factors are significant at the 0.05 level in impacting the probability of someone having heart disease if all else is held constant, include age, sex, systolic blood pressure (SBP), cholesterol level, and self-reported number of cigarettes smoked each week. If two people had the exact same statistics but one person was one year older than that person would be 5.664% more likely to have heart disease than the younger person. If a man and woman had the exact same statistics, the man would be 98.987% more likely to have heart disease. Similarly, holding all other factors constant, an increase SBP increases the chance of heart disease by 1.696%, and 0.448% for an increase in cholesterol level.

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

```
predict(fit4.2, hd_data.new, type="response")
```

```
##    1407
## 0.0336
```

The probability that Liz will have heart disease according to the model is 3.36%.

2.2 Classification analysis

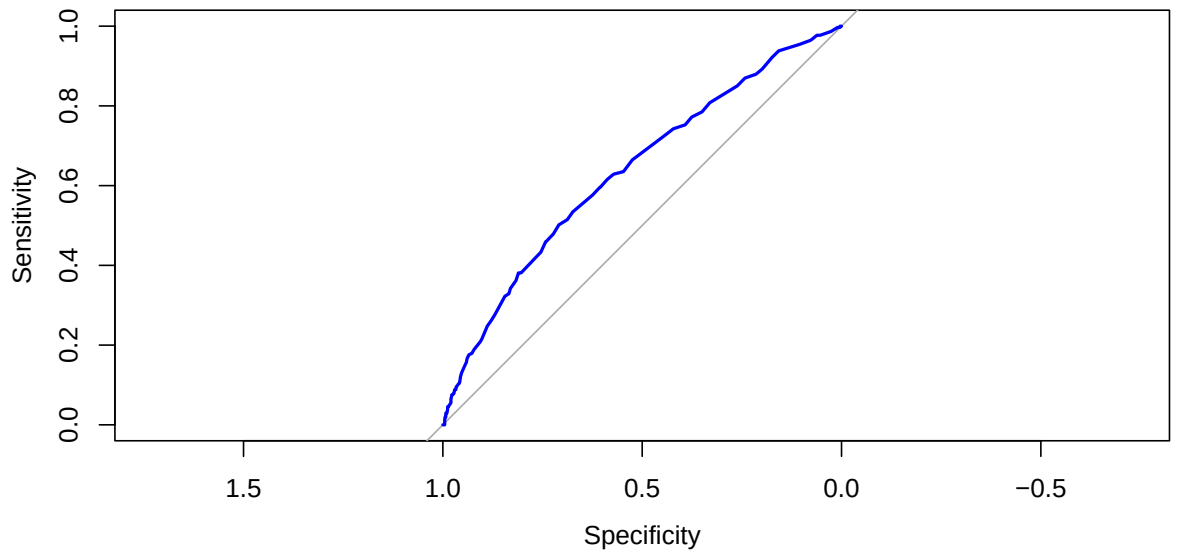
2.2.1 ROC/FDR

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.

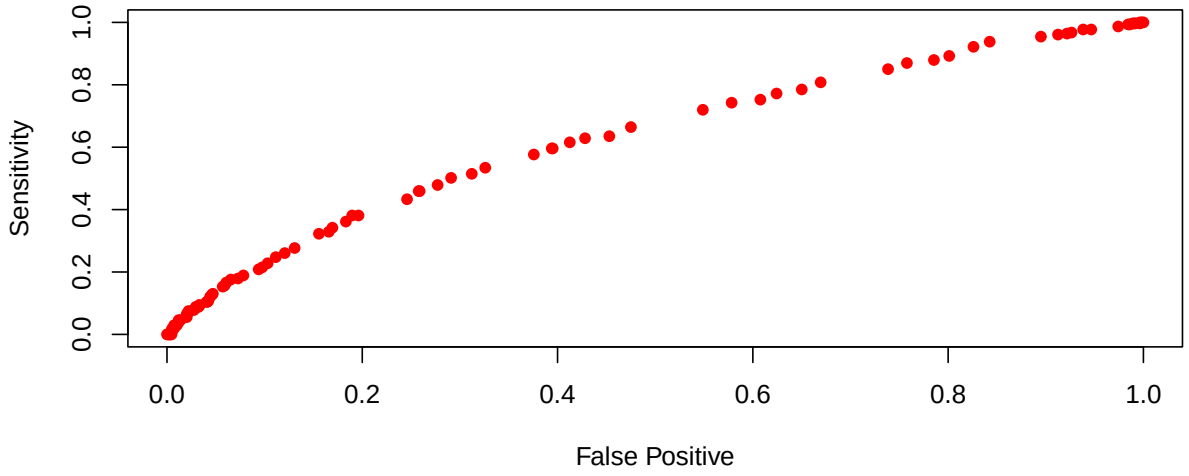
```
fit1.roc<- roc(hd_data.f$HD, fit1$fitted, plot=T, col="blue")
```

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

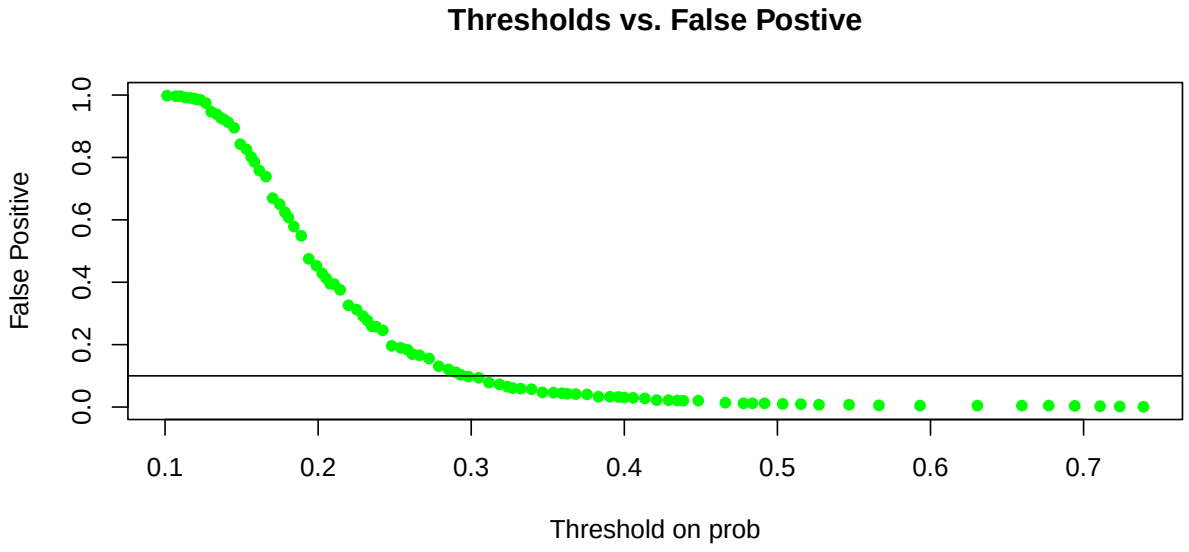
```
## Setting direction: controls < cases
```




```
plot(1-fit1.roc$specificities, fit1.roc$sensitivities, col="red", pch=16,
     xlab="False Positive",
     ylab="Sensitivity")
```




```
plot(fit1.roc$thresholds, 1-fit1.roc$specificities, col="green", pch=16,
     xlab="Threshold on prob",
     ylab="False Positive",
     main = "Thresholds vs. False Postive")
abline(h = 0.1, col = "black")
```



The ROC curve plots pairs of sensitivity and specificity and is useful in selecting a classifier. We want both a high sensitivity and high specificity and aim for a balance of the two. The classifier for a high true positive rate while maintaining a false positive rate less than 0.1 is one with a threshold on probability of around 0.3.

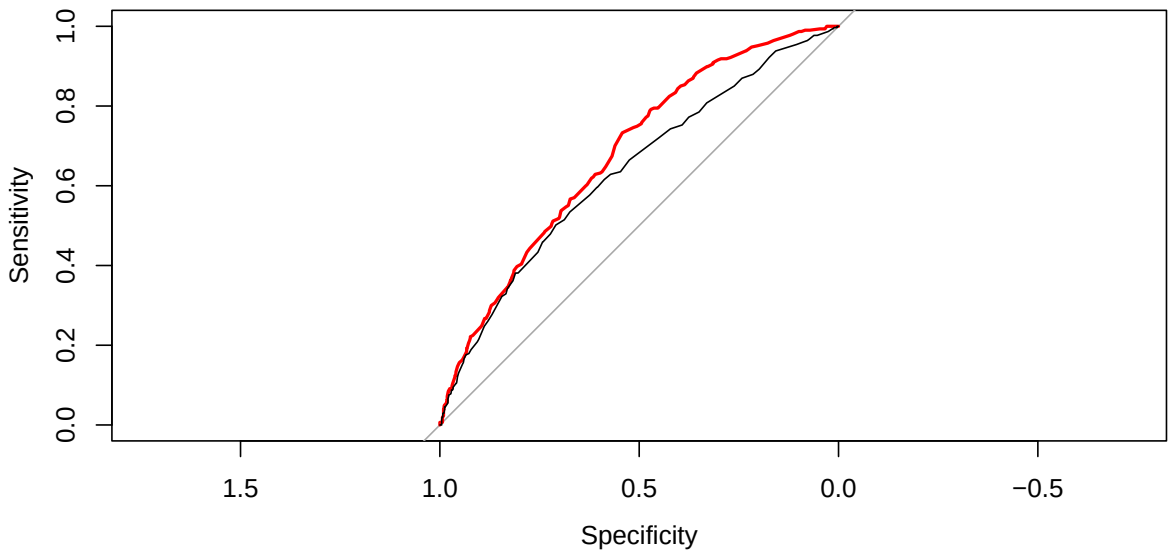
ii. Overlay two ROC curves: one from `fit1`, the other from `fit2`. Does one curve always contain the other curve? Is the AUC of one curve always larger than the AUC of the other one? Why or why not?

```
fit2.roc<- roc(hd_data.f$HD, fit2$fitted, plot=T, col="red")
```

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

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

```
lines(fit1.roc$specificities, fit1.roc$sensitivities)
```




```
fit1.roc$auc
```

```
## Area under the curve: 0.636
```

```
fit2.roc$auc
```

```
## Area under the curve: 0.68
```

The red curve for fit2 always contains the black curve for fit1. The AUC for fit2 is larger at 0.68 than the AUC for fit1 at 0.636. This is because fit2 has an additional variable so it explains more of the response variable.

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?

```
fit1.pred.5 <- ifelse(fit1$fitted > 1/2, "1", "0")
cm.5.1 <- table(fit1.pred.5, hd_data.f$HD)
positive.pred.1 <- cm.5.1[2,2] / sum(cm.5.1[2,])
positive.pred.1
```

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

```
negative.pred.1 <- cm.5.1[1,1] / sum(cm.5.1[1,])
negative.pred.1
```

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

```
fit2.pred.5 <- ifelse(fit2$fitted > 1/2, "1", "0")
cm.5.2 <- table(fit2.pred.5, hd_data.f$HD)
positive.pred.2 <- cm.5.2[2,2] / sum(cm.5.2[2,])
positive.pred.2
```

```
## [1] 0.472
```

```
negative.pred.2 <- cm.5.2[1,1] / sum(cm.5.2[1,])
negative.pred.2
```

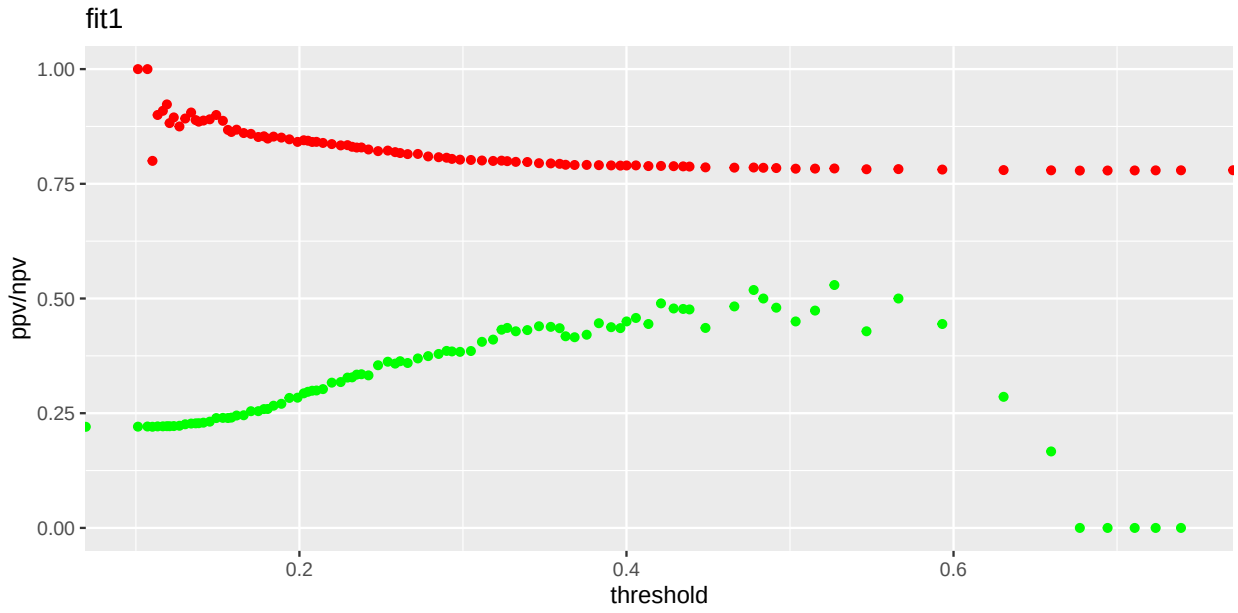
```
## [1] 0.786
```

For fit1, the positive prediction value is 0.45 and the negative prediction value is 0.783. Meanwhile, for fit 2 the positive prediction value is 0.472 and the negative prediction value is 0.786. Fit2 has a higher positive prediction value and is more desirable if prioritizing positive prediction value.

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.

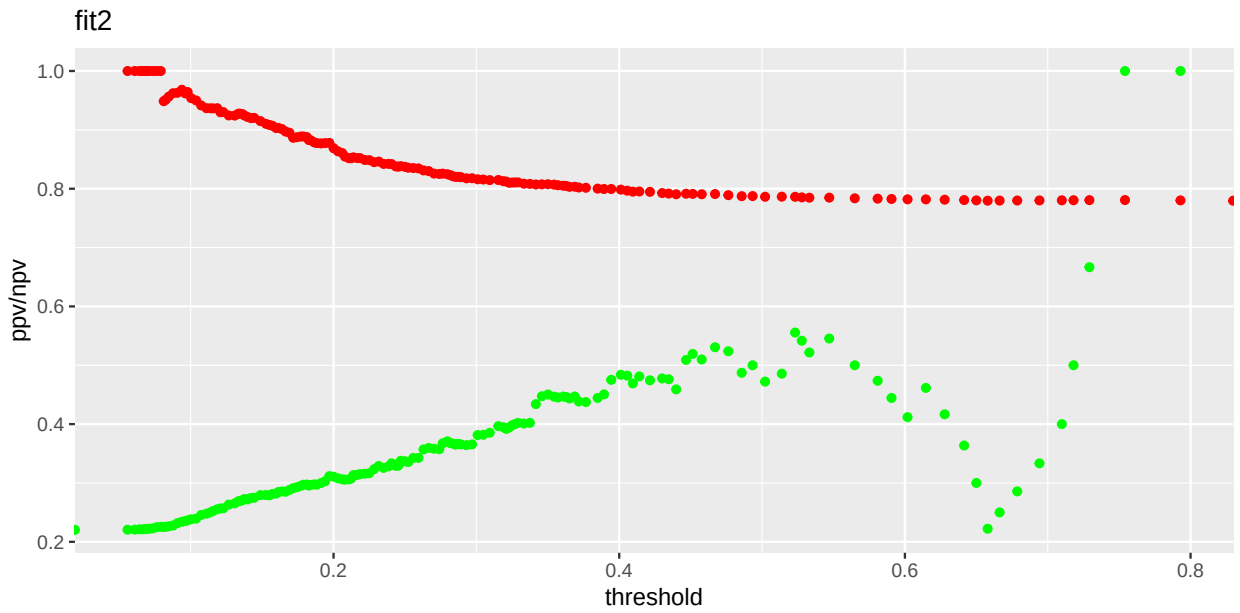
```
coordinates.1.ppv <- coords(fit1.roc, x = "all", input = "threshold", ret = c("threshold", "ppv"))
coordinates.1.npv <- coords(fit1.roc, x = "all", input = "threshold", ret = c("threshold", "npv"))
plot1 <- ggplot() + geom_point(data=coordinates.1.ppv, aes(x=threshold, y=ppv), colour="green") + geom_
plot1
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
```



```
coordinates.2.ppv <- coords(fit2.roc, x = "all", input = "threshold", ret = c("threshold", "ppv"))
coordinates.2.npv <- coords(fit2.roc, x = "all", input = "threshold", ret = c("threshold", "npv"))
plot2 <- ggplot() + geom_point(data=coordinates.2.ppv, aes(x=threshold, y=ppv), colour="green") + geom_
plot2
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
```



We would choose model 2 because the set of points in model 2 have higher ppv and npv values than that of model 1.

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

$a_{01}/a_{10} = \frac{1}{10}\ P(HD = 1 \mid AGE \cap SEX \cap SBP \cap CHOL) > \frac{\frac{1}{10}}{1+\frac{1}{10}}\ P(HD = 1 \mid AGE \cap SEX \cap SBP \cap CHOL) > 0.0909\ logit > log\frac{.0909}{1-.0909}\ logit > -2.303$

The linear boundary is $ - 8.40872+0.05664AGE+0.98987SEXMALE+0.01696SBP+0.00448CHOL > -2.303$

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

```
fit4.2.bayes <- as.factor(ifelse(fit4.2$fitted > 0.0909, "1", "0"))
MCE.final.bayes <- (10*sum(fit4.2.bayes[hd_data.f$HD == "1"] != "1")
              + sum(fit4.2.bayes[hd_data.f$HD == "0"] != "0"))/length(hd_data.f$HD)
MCE.final.bayes
```

```
## [1] 0.719
```

Our estimated weighted misclassification error is 0.719.

iii. How would you classify Liz under this classifier?

$ - 8.40872+0.05664AGE+0.98987SEXMALE+0.01696SBP+0.00448CHOL >? - 2.303$ $ - 8.40872+0.05664(50)+0.98987(0)- = - 2.905$

We would classify Liz as not having heart disease because her predicted logit value of -2.905 is less than the threshold of -2.303.

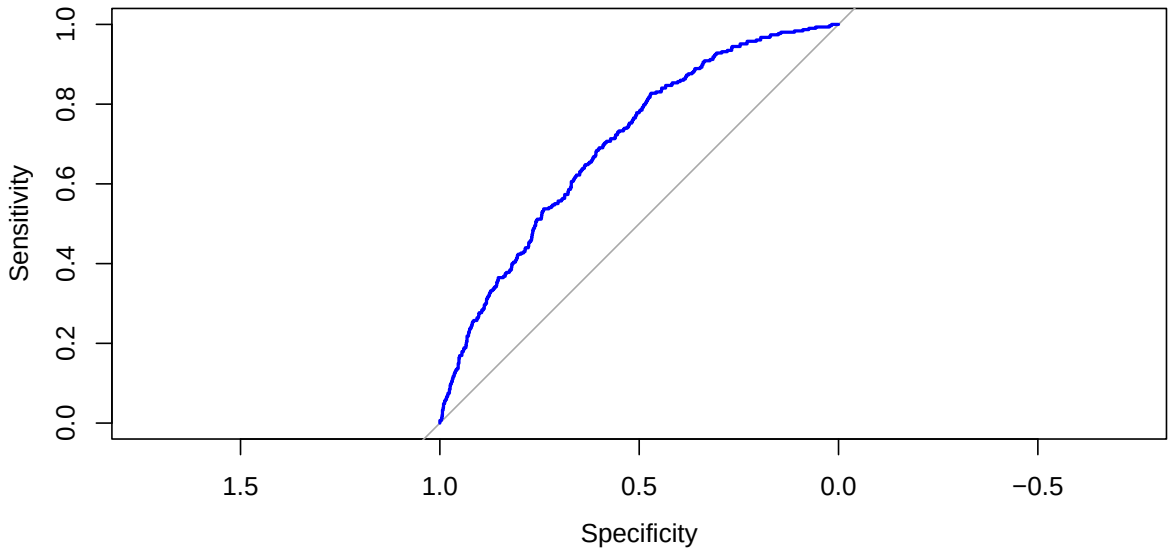
iv. Bayes rule gives us the best rule if we can estimate the probability of `HD-1` accurately. In practice we use logistic regression as our working model. How well does the Bayes rule work in practice? We hope to show in this example it works pretty well.

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

```
fit4.2.roc<- roc(hd_data.f$HD, fit4.2$fitted, plot=T, col="blue")
```

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

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

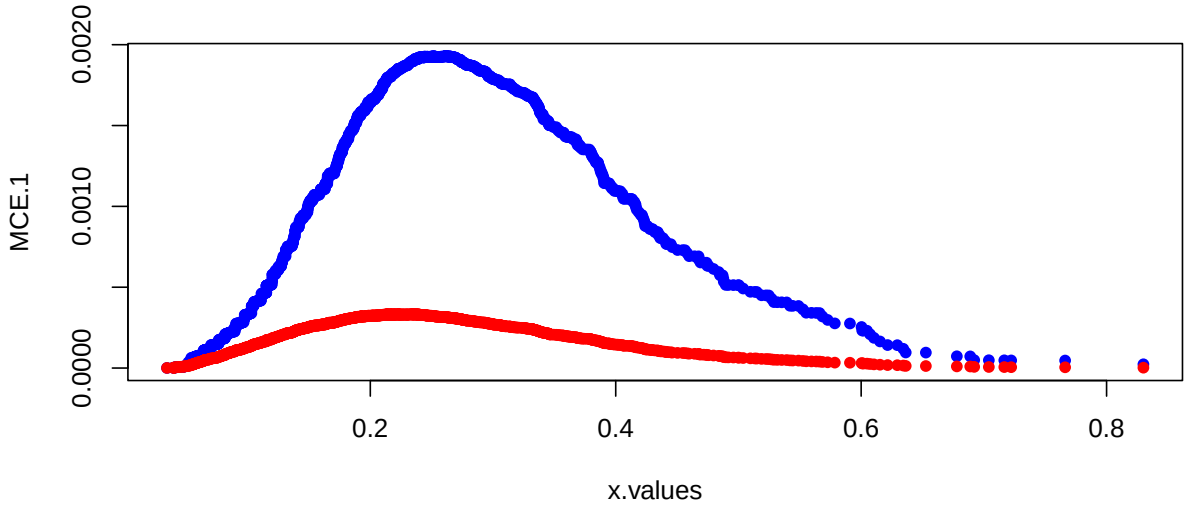



```
x.values <- fit4.2.roc$thresholds

n_s1<-num(fit4.2.roc$sensitivities)
n_s2<-num(fit4.2.roc$specificities)
n_hdf<-nrow(hd_data.f)

MCE.1 <- (10*(1-fit4.2.roc$sensitivities)*n_s1+(1-fit4.2.roc$specificities)*n_s2)/n_hdf
MCE.2 <- ((1-fit4.2.roc$sensitivities)*n_s1+(1-fit4.2.roc$specificities)*n_s2)/n_hdf
```

```
plot(x.values, MCE.1, col="blue", pch=16)
points(x.values, MCE.2, col="red", pch=16)
```



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

The Bayes rule classifier has an increasing MCE with from thresholds 0-0.3 but then decreases after that. The weighted misclassification error is 0.719

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

```
fit4.2.bayes.1 <- as.factor(ifelse(fit4.2$fitted > 0.5, "1", "0"))
MCE.final.bayes.1 <- (sum(fit4.2.bayes.1[hd_data.f$HD == "1"] != "1")
              + sum(fit4.2.bayes.1[hd_data.f$HD == "0"] != "0"))/length(hd_data.f$HD)
MCE.final.bayes.1
```

```
## [1] 0.22
```

The Bayes rule classifier has an increasing MCE with from thresholds 0-0.2 but then decreases after that. The increase and decrease is not as large as if $a_{10}/a_{01} = 10$. The weighted misclassification error is 0.22.

3 Part II: Project

3.1 Project Option 1 Credit Risk via LendingClub

3.2 Project Opetion 2 Diabetes and Health Management