Modern Data Mining, HW 3

Group Member Mahika Calyanakoti Group Member Graham Branscom
Group Member Andrew Raines

11:59 pm, 03/17, 2024

Contents

1 PartI: Model Building

Multiple regression is one of the most popular methods used in statistics as well as in machine learning. We use linear models as a working model for its simplicity and interpretability. It is important that we use domain knowledge as much as we could to determine the form of the response as well as the function format for the factors. Then, when we have many possible features to be included in the working model it is inevitable that we need to choose a best possible model with a sensible criterion. Regularizations such as LASSO are introduced. Be aware that if a model selection is done formally or informally, the inferences obtained with the final `lm()` fit may not be valid. Some adjustment will be needed. This last step is beyond the scope of this class. Check the research line that Linda and collaborators have been working on.

The main job in this part is a rather involved case study about devastating covid19 pandemic. Please read through the case study first. This project is for sure a great one listed in your CV.

For covid case study, the major time and effort would be needed in EDA portion.

1.1 Objectives

- Model building process
- Methods
  - Model selection
    - LASSO (L1 penalty)
    - A quick backward elimination
- Understand the criteria
  - Testing Errors
  - `K fold Cross Validation`
  - `LASSO`
- Packages
  - `lm()`, `Anova`
  - `regsubsets()`
  - `glmnet()` & `cv.glmnet()`

1.2 Review materials

- Study lecture: Regularization
- Study lecture: Multiple regression

Review the code and concepts covered during lectures: multiple regression, penalized regression through elastic net (only LSASSO).

Important Notice: The focus of this part is Covid case study.

1.3 Case study: COVID19

See a seperate file covid_case_study_2024.Rmd for details.

- Start the EDA as earlier as possible.
- Please check previous midterms where we used the same dataset.

2 Part II: Logistic Regression

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. We will not use it here in this work.

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
- R functions/Packages
  - `glm()`, `Anova`
  - `pROC`
- Data needed
  - `Framingham.dat`

2.2 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
```

2.2.1 Identify risk factors

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

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

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

The likelihood function = Prob(All events in data occurred)=P((y1 = 1, x1 = 140) ^ (y2 = 0, x2 = 110) ^ (y3 = 1, x3 = 150) ^ (y4 = 1, x4 = 260) ^ (y5 = 0, x5 = 122)) Since these are independent observations, we can multiply them: =P(y1 = 1, x1 = 140) * P(y2 = 0, x2 = 110) * P(y3 = 1, x3 = 150) * P(y4 = 1, x4 = 260) * P(y5 = 0, x5 = 122)

$$lik(\beta_0, \beta_1|data) = \frac{e^{\beta_0+\beta_1 140}}{1 + e^{\beta_0+\beta_1 140}} \times \frac{1}{1 + e^{\beta_0+\beta_1 110}} \times \frac{e^{\beta_0+\beta_1 150}}{1 + e^{\beta_0+\beta_1 150}} \times \frac{e^{\beta_0+\beta_1 260}}{1 + e^{\beta_0+\beta_1 260}} \times \frac{1}{1 + e^{\beta_0+\beta_1 122}}$$

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 is obtained by finding a B1 and B0 that maximize the likelihood function above:

$$(\hat{\beta}_1, \hat{\beta}_0) = \arg\max_{\beta_0,\beta_1} \mathcal{L}(\beta_0, \beta_1|Data)$$

Look at the fit1 applied to the five random samples from above:

- logit = -3.66 + 0.0159 SBP
-

$$\hat{P}(HD = 1|SBP) = \frac{e^{-3.66+0.0159\times SBP}}{1 + e^{-3.66+0.0159\times SBP}}$$
$$\hat{P}(HD = 0|SBP) = \frac{1}{1 + e^{-3.66+0.0159\times SBP}}$$

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`. Since our logit function is only based on SBP, we can calculate her probability of having HD as follows:

Based on fit1 we plug in `SBP` value into the prob equation.

$$\hat{P}(HD = 1|SBP = 110) = \frac{e^{-3.66+0.0159\times SBP}}{1 + e^{-3.66+0.0159\times SBP}} = \frac{e^{-3.66+0.0159\times 110}}{1 + e^{-3.66+0.0159\times 110}} \approx 0.128$$

We can also use the `predict()` function. We see that we also get 0.128.

```
##     1
## 0.128
```

2.2.1.2 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`.

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.

We added sex because according to our final model from above (fit1.6), sex has the lowest p-value.

```
        Estimate Std..Error z.value Pr...z..
```

(Intercept) -7.3775 0.82488 -8.94 3.76e-19 SBP 0.0174 0.00235 7.39 1.49e-13 AGE 0.0570 0.01440 3.96 7.65e-05 SEXMALE 0.9012 0.14081 6.40 1.55e-10

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

The residual deviance for fit1 was 1417.5; the residual deviance for fit2 was 1357.9. The residual deviance will always decrease as the number of variables included increases, but it is not always the best idea to include all of the variables since the data will be "overfitted" by the model.

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 do a Wald test. We can see that Sex is significant at alpha=0.01 since it has a p-value of 1.6e-10.

From anova, we can see that fit1 and fit2 are significantly different.

Anova on fit2 tells us that the p-value for sex is significant at alpha=0.01 (6.0e-11), which is smaller than the p-value found from the Wald test (1.6e-10).

2.2.1.3 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.

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?

We use bestglm() to find the model with the smallest AIC:

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

```
## [1] "BestModel"   "BestModels"  "Bestq"       "qTable"      "Subsets"
## [6] "Title"       "ModelReport"
```

List the top 5 models. In the way any one of the following model could be used.

```
##    AGE SEXFEMALE SEXMALE  SBP   DBP CHOL   FRW   CIG Criterion
## 1 TRUE     FALSE    TRUE TRUE FALSE TRUE  TRUE  TRUE      1355
## 2 TRUE      TRUE   FALSE TRUE FALSE TRUE  TRUE  TRUE      1355
## 3 TRUE     FALSE    TRUE TRUE FALSE TRUE FALSE  TRUE      1356
## 4 TRUE      TRUE   FALSE TRUE FALSE TRUE FALSE  TRUE      1356
## 5 TRUE     FALSE    TRUE TRUE FALSE TRUE FALSE FALSE      1357
```

We run a summary and Anova based on the first row from BestModels, which has the lowest AIC value. We can see that our exhaustive search does NOT guarantee that the p-values for all the remaining variables are

less than .05. For instance, FRW has p-value of 0.1315 from the Wald's test and 0.1336 from the Chi-squared test.

Our final model from the exhaustive search is not the same as that from our backwards elimination from above since here we included two extra variables: FRW and CIG.

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

In the context of the final model from the exhaustive search, "important factors" refer to the variables (AGE, SEX, SBP, CHOL, FRW, CIG) and their coefficients that are statistically significant in predicting the likelihood of heart disease. These variables are all positively correlated with heart disease since they have positive B-values. Overall, all are statistically significant at alpha = 0.05, except FRW. We see a low AIC of 1357, which suggests our model is parsimonious.

AGE (0.06153): For each additional year of age, the log odds of having heart disease increase by 0.06153, assuming all other variables are held constant.

SEX (0.91127 for male): Being male (SEXMALE) increases the log odds of having heart disease by 0.91127 compared to being female, assuming all other variables are held constant.

SBP (0.01597): For each additional unit increase in systolic blood pressure (SBP), the log odds of having heart disease increase by 0.01597, assuming all other variables are held constant.

CHOL (0.00449): For each additional unit increase in cholesterol (CHOL) level, the log odds of having heart disease increase by 0.00449, assuming all other variables are held constant.

FRW (0.00604): The log odds of having heart disease increase by 0.00604 for each unit increase in FRW, assuming all other variables are held constant. However, FRW is not statistically significant at an alpha = 0.05 (FRW p-value = 0.1315).

CIG (0.01228): For each additional unit increase in cigarette consumption (CIG), the log odds of having heart disease increase by 0.01228, assuming all other variables are held constant. CIG is marginally significant but still under the alpha = 0.05 level (p-value = 0.0437).

From this summary, we can conclude that age, sex (male), systolic blood pressure (SBP), and cholesterol (CHOL) are important factors significantly associated with the likelihood of heart disease, based on our final model.

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

We can also use the `predict()` function. We see that her probability of having HD is 0.0459. This is lower from our previous prediction of 0.128.

```
##      1
## 0.0496
```

2.2.2 Classification analysis

2.2.2.1 ROC/FDR

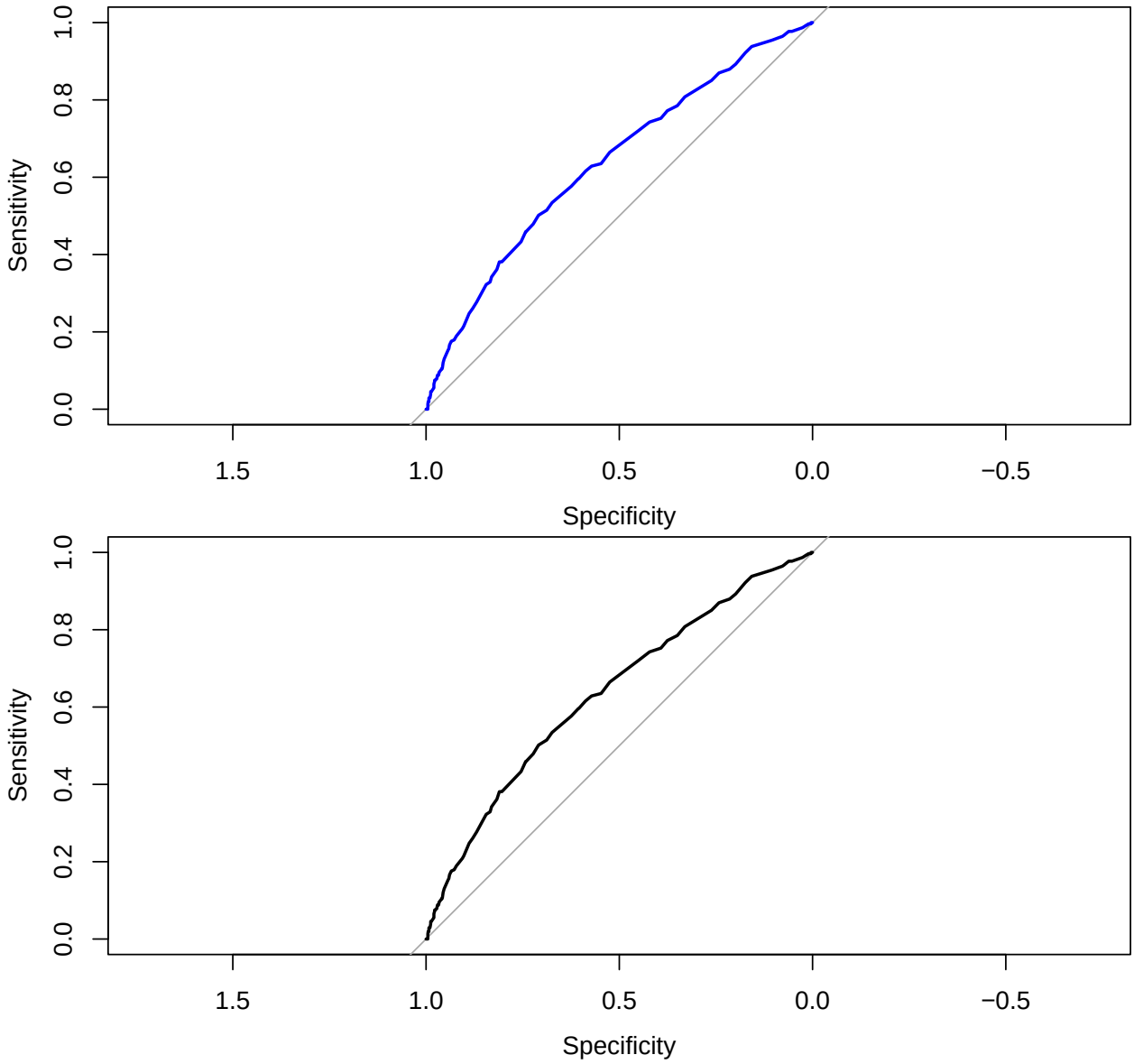
i. Display the ROC curve using `fit1`. Explain what ROC reports and how to use the graph. Specify the classifier such that the False Positive rate is less than .1 and the True Positive rate is as high as possible.

Sensitivity = TP rate Specificity = 1 - FP

We want FPR to be < 0.1, so we want want specificity to be > 0.9.

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

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



We can find the thresholds from `fit1.roc`.

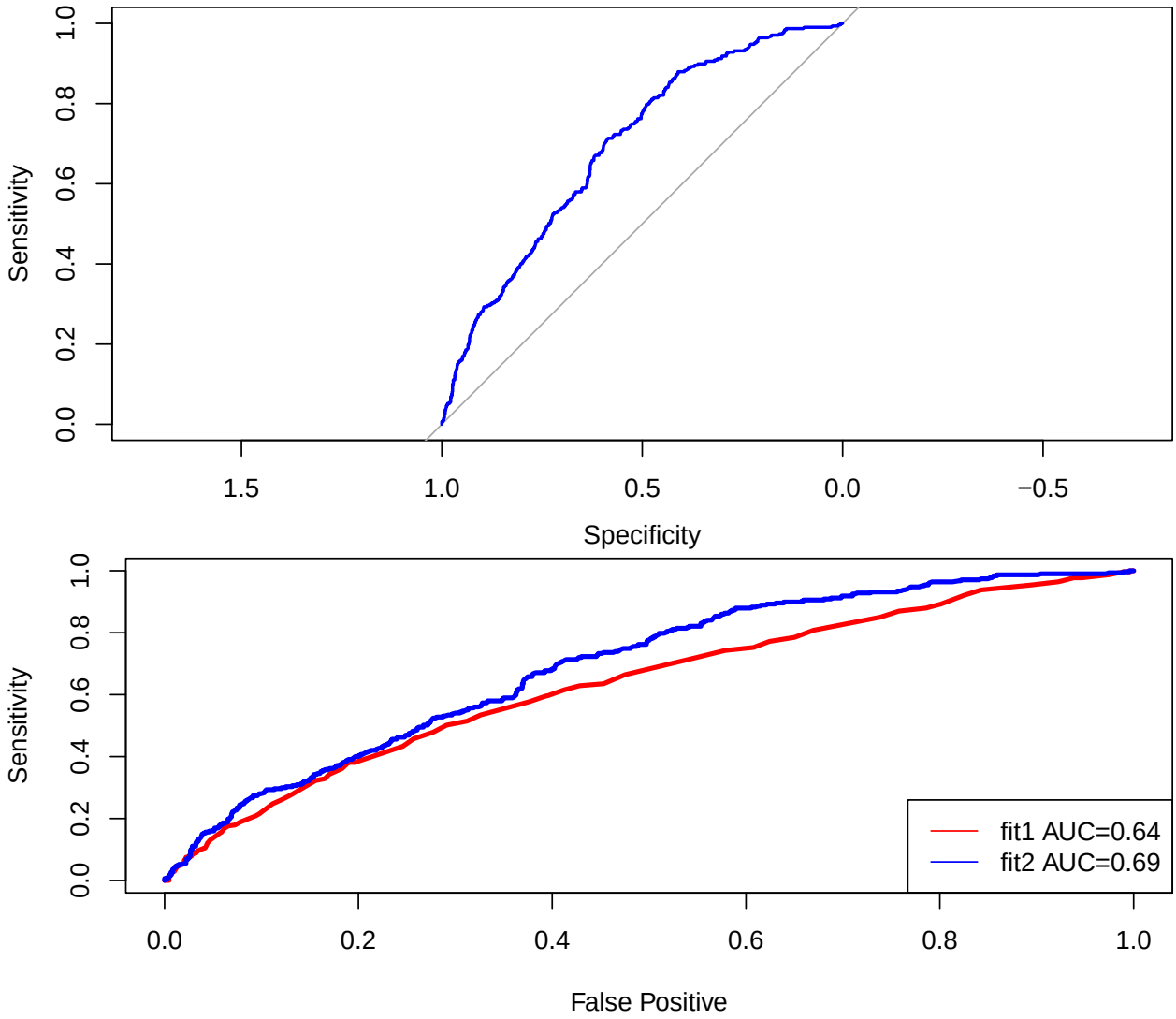
We find the corresponding specificity (0.902) and sensitivity (0.215) at the optimal_threshold (0.298).

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?

When overlaying ROC curves from the two classifiers, one curve doesn't necessarily always contain the other, and the AUC of one curve isn't always larger than the other. The ROC curves represent the trade-off between true positive rate and false positive rate at different classification thresholds, which can vary between classifiers. Thus, their overlap, containment, or relative AUC values depend on the specific performance characteristics and threshold selections of the classifiers being compared. In this specific case, however, one curve does contain the other (fit2 contains fit1).

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

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



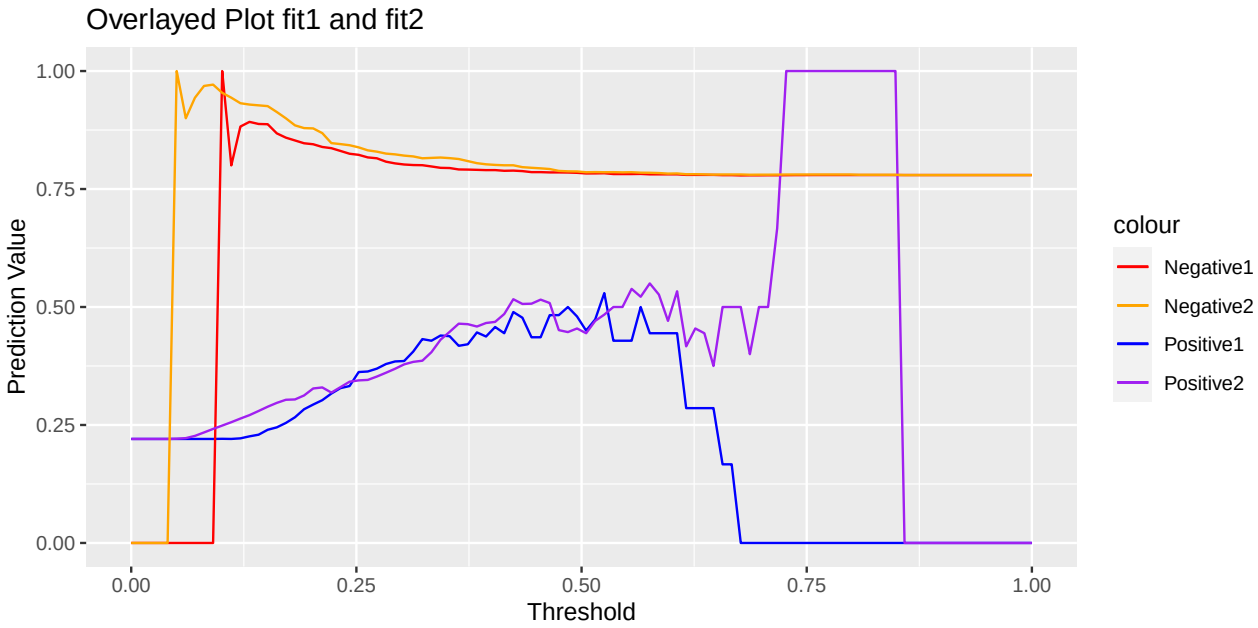
iii. Estimate the Positive Prediction Values and Negative Prediction Values for `fit1` and `fit2` using .5 as a threshold. Which model is more desirable if we prioritize the Positive Prediction values?

For fit1: Positive Prediction Value = 0.45 Negative Prediction Value = 0.783

For fit2: Positive Prediction Value = 0.415 Negative Prediction Value = 0.786

Thus, fit1 prioritizes the Positive Prediction values. This is different than if we were to use AUC, given that fit2 has a higher AUC than fit1.

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.



fit2's curves are consistently higher compared to fit1's curves. If the set of positive and negative prediction values are the concerns then we would choose fit2 as it is more robust wrt threshold perturbations.

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

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.

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?