STOR767 - HW 2 Computing

Leo Li (PID: 730031954)

Problem: Logistic Regression and Classification

We will continue to use the Framingham Data (`Framingham.dat`) so that you are already familiar with the data and the variables. To keep our answers consistent, 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 307 people diagnosed with heart disease and 1086 without heart disease.

```
   0    1
1086  307
```

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

```
summary(hd_data.f)
```

```
##  HD            AGE            SEX           SBP             DBP
##  0:1086   Min.   :45.00   FEMALE:730   Min.   : 90.0   Min.   : 50.00
##  1: 307   1st Qu.:48.00   MALE  :663   1st Qu.:130.0   1st Qu.: 80.00
##           Median :52.00                Median :142.0   Median : 90.00
##           Mean   :52.43                Mean   :148.1   Mean   : 90.16
##           3rd Qu.:56.00                3rd Qu.:160.0   3rd Qu.: 98.00
##           Max.   :62.00                Max.   :300.0   Max.   :160.00
##       CHOL            FRW             CIG
##  Min.   : 96.0   Min.   : 52.0   Min.   : 0.000
##  1st Qu.:200.0   1st Qu.: 94.0   1st Qu.: 0.000
##  Median :230.0   Median :103.0   Median : 0.000
##  Mean   :234.6   Mean   :105.4   Mean   : 8.035
##  3rd Qu.:264.0   3rd Qu.:114.0   3rd Qu.:20.000
##  Max.   :430.0   Max.   :222.0   Max.   :60.000
```

A) Create a traning dataset with 1000 observations and a testing dataset with the rest of the data. Using `set.seed(1)`.

The following program is used to create a training dataset with 1000 observations and a testing dataset with the rest of the data. The traing data is stored in `hd_data.train`, and the testing dataset is stored in `hd_data.test`.

```
N <- length(hd_data.f$HD)
set.seed(1)
index.train <- sample(N, 1000)
hd_data.train <- hd_data.f[index.train,]
hd_data.test <- hd_data.f[-index.train,]
```

Now, we can check the dimension of the training dataset:

```r
dim(hd_data.train)
```

```
## [1] 1000    8
```

We can also check the dimension of the testing dataset:

```r
dim(hd_data.test)
```

```
## [1] 393   8
```

B) Our goal is to fit a well-fitting model, that is still small and easy to interpret (parsimonious).

1. Use AIC as the criterion for model selection. Find a logistic regression model with small AIC through exhaustive search in the training dataset. Call this model `fit.aic`.

In this question, we use AIC as the criterion for model selection, and find a model with small AIC through exhaustive search in the training dataset.

```r
qb1 <- hd_data.train[, c("AGE","SEX","SBP","DBP","CHOL","FRW","CIG","HD")]
fit.aic = bestglm(qb1, family=binomial, IC="AIC", method = "exhaustive")
```

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

```r
fit.aic
```

```
## AIC
## BICq equivalent for q in (0.88394303377024, 0.96785765478523)
## Best Model:
##                  Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -10.013905633 1.204649820 -8.312711 9.354024e-17
## AGE           0.070970307 0.017748842  3.998588 6.372148e-05
## SEXMALE       0.841154718 0.183481432  4.584413 4.552629e-06
## SBP           0.015800228 0.003031632  5.211789 1.870278e-07
## CHOL          0.004877991 0.001782869  2.736035 6.218443e-03
## FRW           0.008489719 0.004717122  1.799767 7.189748e-02
## CIG           0.012740214 0.007239669  1.759778 7.844537e-02
```

2. Use the model chosen from part B1 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). Make sure to define "important factors" in your words.

We choose the model from part B1 as our final model. That is, our final model is selected as the one with the lowest AIC value from the exhaustive search among the possible logistic regression models that we can construct from the Framingham Study dataset.

Here, we would define the "important factors" as the factors in our final model with p-values less than 0.05, which consist of AGE, SEX, SBP, and CHOL. We can describe the relationships between those variables and heart disease as the following:

- AGE: While controlling for all the other important factors as well as FRW and CIG, if there is one unit increase in AGE, then the log of the odds of getting heart disease is expected to increase by 0.071;
- SEX: While controlling for all the other important factors as well as FRW and CIG, the log of the odds ratio of getting heart disease for males compared to females is expected to be 0.841;
- SBP: While controlling for all the other important factors as well as FRW and CIG, if there is one unit increase in SBP, then the log of the odds of getting heart disease is expected to increase by 0.016;
- CHOL: While controlling for all the other important factors as well as FRW and CIG, if there is one unit increase in CHOL, then the log of the odds of getting heart disease is expected to increase by 0.005;

3. Liz is a patient with the following readings: `AGE=50, GENDER=FEMALE, SBP=110, DBP=80, CHOL=180, FRW=105, CIG=0`. What is the probability that she will have heart disease, according to our final model?

According to our final model, we can estimate Liz's probability of getting heart disease as:

$$\frac{1}{1+\exp[-(-10.014+50\times 0.071+110\times 0.016+180\times 0.005+105\times 0.008)]}=0.049$$

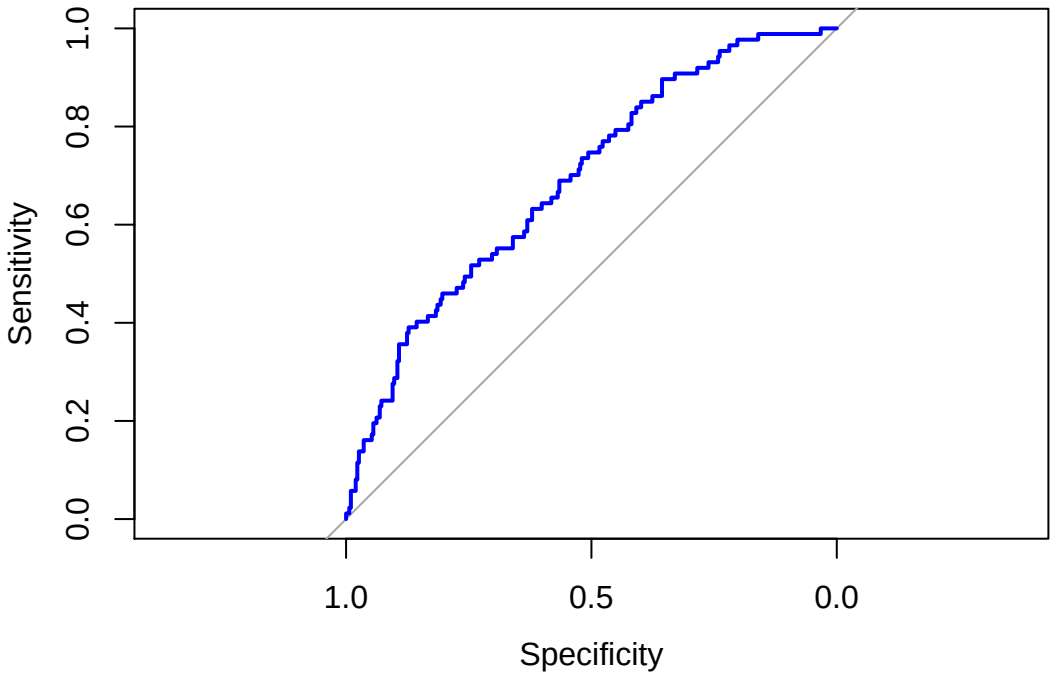
4. Consider using `fit.aic` for classification in the test dataset. Display the ROC curve using `fit.aic`. Explain what ROC reports and how to use the graph.

Here, we use `fit.aic` for classification in the test dataset, and the ROC curve using `fit.aic` is present as the following:

```
fit1 <- glm(HD~AGE+SEX+SBP+CHOL+FRW+CIG, hd_data.train, family=binomial(logit))
fit1.fitted.test <- predict(fit1, hd_data.test, type="response")
fit1.roc<- roc(hd_data.test$HD, fit1.fitted.test, plot=T, col="blue")
```

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

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



In the context of this question, we can explain the ROC curve as the following:

In logistic regression, as what has been done in this question, we often output a probability,

$$P(y=1|X=x)$$

Then, by changing the thresholds, we will obtain a family of classifiers, and we may plot all the pairs of false positives on the x-axis and true positive on the y-axis to have the resulting ROC curve.

The ROC is used to evaluate the performance of a classifier. In general, we would prefer an ROC curve that is closer to the upper left coner of the graph, or in other words, an ROC curve that has greater area under the curve (AUC).

C) 1. Use BIC as the criterion for model selection. Find a logistic regression model with small BIC through exhaustive search. Call this model `fit.bic`. Compare `fit.bic` and `fit.aic`.

Now, we use BIC as the criterion for model selection. In order to be comparable to part B1, we also conduct the analysis on the training dataset. Through exhaustive search, a logistic regression model with small BIC is presented as the following:

```
qb1 <- hd_data.train[, c("AGE","SEX","SBP","DBP","CHOL","FRW","CIG","HD")]
fit.bic = bestglm(qb1, family=binomial, IC="BIC", method = "exhaustive")
```

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

```
fit.bic
```

```
## BIC
## BICq equivalent for q in (0.432682768631664, 0.88394303377024)
## Best Model:
##                 Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -8.936468575 1.088534826 -8.209630 2.218704e-16
## AGE          0.065597502 0.017409542  3.767905 1.646233e-04
## SEXMALE      0.905415959 0.169628542  5.337639 9.416460e-08
## SBP          0.017077991 0.002892304  5.904632 3.534349e-09
## CHOL         0.004848089 0.001772726  2.734821 6.241413e-03
```

Comparing the above results to the model results of `fit.aic` presented in part B1, we can realize that the model `fit.bic` is more parsimonious than `fit.aic`.

2. Overlay two ROC curves with the test dataset: One from `fit.bic`, the other from `fit.aic` from part A1. Based on the ROC curves, which one do you prefer?

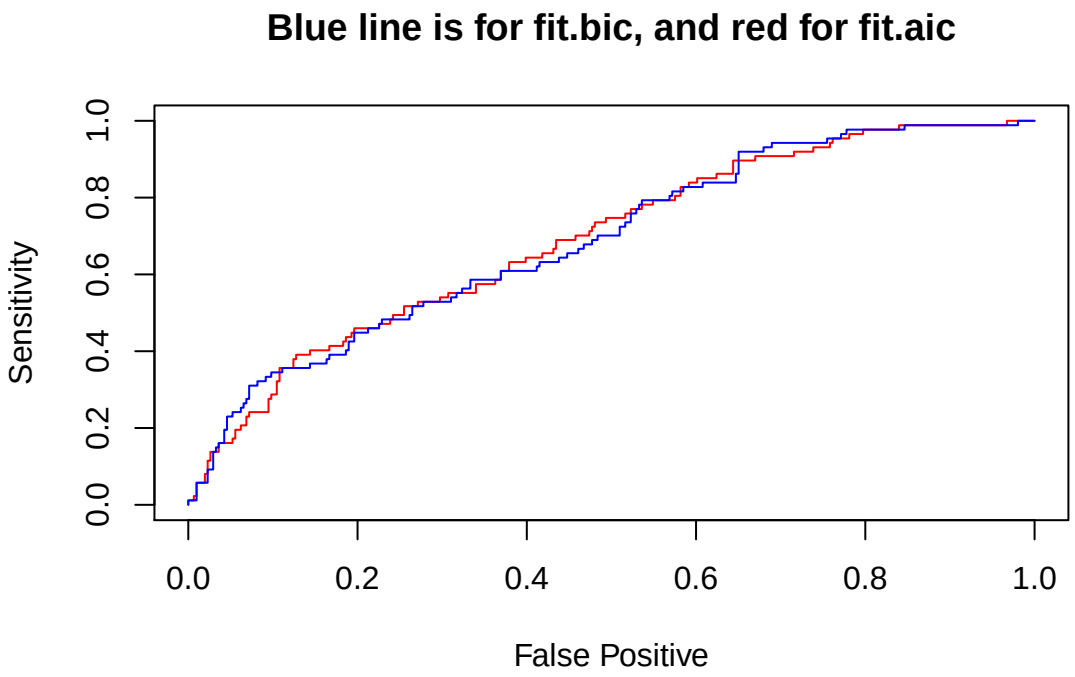
Now, we overlay the two ROC curves with the test dataset:

```
fit2 <- glm(HD~AGE+SEX+SBP+CHOL, hd_data.train, family=binomial(logit))
fit2.fitted.test <- predict(fit2, hd_data.test, type="response")
fit2.roc<- roc(hd_data.test$HD, fit2.fitted.test, plot=F, col="blue")
```

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

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

```
plot(1-fit1.roc$specificities, fit1.roc$sensitivities, col="red", pch=16, cex=.7,type="l",
     xlab="False Positive",
     ylab="Sensitivity")
points(1-fit2.roc$specificities, fit2.roc$sensitivities, col="blue", type="l",pch=16, cex=.6)
title("Blue line is for fit.bic, and red for fit.aic")
```



Now, we consider the AUC for `fit.bic`:

```
pROC::auc(fit2.roc)
```

```
## Area under the curve: 0.6886
```

And also the AUC for `fit.aic`:

```
pROC::auc(fit1.roc)
```

```
## Area under the curve: 0.6904
```

According to the ROC curves as well as the corresponding AUC, we can realize that `fit.aic` and `fit.bic` has very familiar performance. However, `fit.bic` is more parsimonious than `fit.aic`, so that I like `fit.bic` better.