STOR767 - Computational Problems for HW 2

Brian N. White

Problem: Logistic Regression and Classification

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(leaps)
library(bestglm)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

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

Solution

```
set.seed(1)

#number of total observations
N <- length(hd_data.f$HD)
#generate indices for training data
index.train <- sample(N, 1000)
#training data
hd_data.train <- hd_data.f[index.train,]
#testing data
hd_data.test <- hd_data.f[-index.train,]
```

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`.

Solution

An exhaustive search using AIC as the selection criterion returned the logistic regression model with six predictors: AGE, SEX, SBP, CHOL, FRW and CIG. The optimal model, according to this search, has an AIC value of 941.42. Note, I used the package, and corresponding command, 'bestglm' to perform this search.

```
#create data frame of the form Xy where X is the design matrix
#and y is the response vector
df_glm <- data.frame(cbind(hd_data.train[,-1], HD=hd_data.train[,1]))
#use the bestglm package to perform model selection via
#exhaustive search using AIC as the selection criterion
exhaustive_search <- bestglm(df_glm, family = binomial, IC="AIC", method = "exhaustive")
```

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

```r
#returns the best n predictor models where n ranges from 0 to 7
exhaustive_search$Subsets
```

```
##    Intercept   AGE   SEX   SBP   DBP  CHOL   FRW   CIG logLikelihood       AIC
## 0       TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE     -526.9080 1053.8159
## 1       TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE     -507.8731 1017.7462
## 2       TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE     -494.5978  993.1957
## 3       TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE     -487.4175  980.8351
## 4       TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE     -483.6927  975.3855
## 5       TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE     -482.3764  974.7527
## 6*      TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE     -480.8456  973.6912
## 7       TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE     -480.7966  975.5933
```

```r
#the best model over the exhaustive search is the
#model with the 6 predictors specified via the output below
exhaustive_search$BestModel
```

```
##
## Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Coefficients:
## (Intercept)          AGE      SEXMALE          SBP         CHOL          FRW
##  -10.013906     0.070970     0.841155     0.015800     0.004878     0.008490
##         CIG
##    0.012740
##
## Degrees of Freedom: 999 Total (i.e. Null);  993 Residual
## Null Deviance:       1054
## Residual Deviance: 961.7     AIC: 975.7
```

```r
#fit the selected model
fit.aic <- glm(HD ~ AGE + SEX + SBP + CHOL + FRW + CIG, family=binomial, data=hd_data.train)
#AIC is 941.4
summary(fit.aic)
```

```
##
## Call:
## glm(formula = HD ~ AGE + SEX + SBP + CHOL + FRW + CIG, family = binomial,
##     data = hd_data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7439  -0.7234  -0.5477  -0.3344   2.4607
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.013906   1.204650  -8.313  < 2e-16 ***
## AGE           0.070970   0.017749   3.999 6.37e-05 ***
## SEXMALE       0.841155   0.183481   4.584 4.55e-06 ***
## SBP           0.015800   0.003032   5.212 1.87e-07 ***
## CHOL          0.004878   0.001783   2.736  0.00622 **
## FRW           0.008490   0.004717   1.800  0.07190 .
```

```
## CIG           0.012740   0.007240   1.760  0.07845 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1053.82  on 999  degrees of freedom
## Residual deviance:  961.69  on 993  degrees of freedom
## AIC: 975.69
##
## Number of Fisher Scoring iterations: 4
```

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.

Solution

All of the Wald tests, except for FRW, are statistically significant at the $\alpha = 0.05$ level. Thus, the data suggest that there is, in fact, a linear association between the co-variates and the logit of heart disease. To be more precise, each of the variables are positively correlated with the logit of heart disease. Of the co-variates, the sex and, to a much lesser extent, the age of an individual have the greatest impact on the logit of heart disease (e.g. An increase of one year results in a mean increase of 0.94 for the logit of heart disease).

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?

Solution

The probability that Liz will have heart disease, according to the final model, is ~0.04.

```
#create data.frame to store new observation
new_obs <- data.frame(AGE=50, SEX="FEMALE", SBP=110, DBP=80, CHOL=180, FRW=105, CIG=0)
#returns predicted probability of heart disease for new observation
predict(fit.aic, new_obs, type="response")
```

```
##          1
## 0.04936417
```

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.

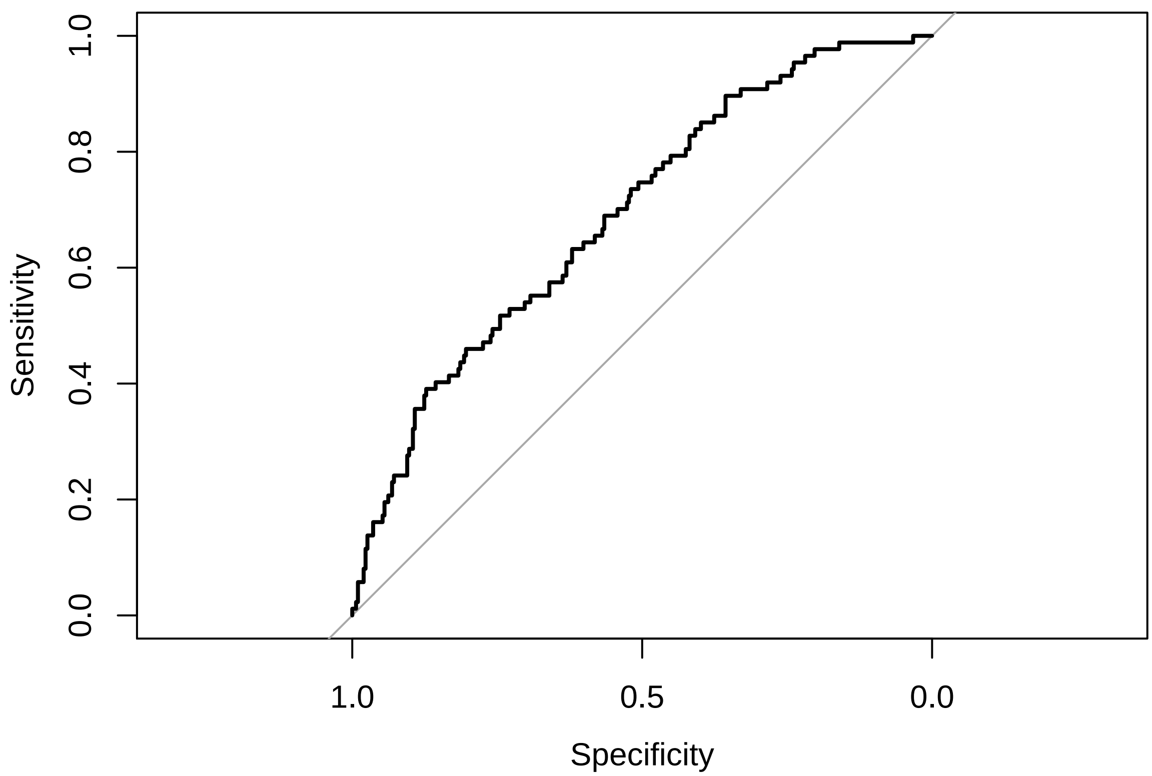
Solution

The ROC curve reports the sensitivity vs specificity for a given classifier as the classification threshold varies. Thus, the ROC curve can be used to measure the performance of a classifier. One metric associated with a ROC curve is the AUC or 'area under curve'. Intuitively, as the AUC increases so to does the corresponding classifier's performance.

```
fit.aic.test <- predict(fit.aic, hd_data.test, type="response")
fit.aic.roc <- roc(hd_data.test$HD, fit.aic.test, plot=T)
```

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

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



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`.

Solution

An exhaustive search using BIC as the selection criterion returned the logistic regression model with four predictors: AGE, SEX, SBP and CHOL. The optimal model, according to this search, has an BIC value of 961.86. (note: The code-chunk below is a modified version of the code-chunk in part B1. Only the information criterion argument, IC, in the bestglm has been changed).

```
#use the bestglm package to perform model selection
#via exhaustive search using BIC as the selection criterion
exhaustive_search2 <- bestglm(df_glm, family = binomial, IC="BIC", method = "exhaustive")
```

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

```
#returns the best n predictor models where n ranges from 0 to 7
exhaustive_search2$Subsets
```

```
##    Intercept   AGE   SEX   SBP   DBP  CHOL   FRW   CIG logLikelihood       BIC
## 0       TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE     -526.9080 1053.8159
```

```
## 1       TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE     -507.8731 1022.6540
## 2       TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE     -494.5978 1003.0112
## 3       TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE     -487.4175  995.5583
## 4*      TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE     -483.6927  995.0165
## 5       TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE     -482.3764  999.2915
## 6       TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE     -480.8456 1003.1377
## 7       TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE     -480.7966 1009.9476
```

```r
#the best model over the exhaustive search is the
#model with the 4 predictors specified via the output below
exhaustive_search2$BestModel
```

```
##
## Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Coefficients:
## (Intercept)          AGE      SEXMALE          SBP         CHOL
##   -8.936469     0.065598     0.905416     0.017078     0.004848
##
## Degrees of Freedom: 999 Total (i.e. Null);  995 Residual
## Null Deviance:       1054
## Residual Deviance: 967.4     AIC: 977.4
```

```r
#fit the selected model
fit.bic <- glm(HD ~ AGE + SEX + SBP + CHOL, family=binomial, data=hd_data.train)
#BIC is 961.8568
summary(fit.bic)
```

```
##
## Call:
## glm(formula = HD ~ AGE + SEX + SBP + CHOL, family = binomial,
##     data = hd_data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6074  -0.7275  -0.5526  -0.3494   2.4449
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.936469   1.088535  -8.210  < 2e-16 ***
## AGE          0.065598   0.017410   3.768 0.000165 ***
## SEXMALE      0.905416   0.169629   5.338 9.42e-08 ***
## SBP          0.017078   0.002892   5.905 3.53e-09 ***
## CHOL         0.004848   0.001773   2.735 0.006241 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1053.82  on 999  degrees of freedom
## Residual deviance:  967.39  on 995  degrees of freedom
## AIC: 977.39
##
## Number of Fisher Scoring iterations: 4
```

Now, let us compare the two models. The most obvious difference is that fit.bic consists of a four variable subset of the predictors in fit.aic. All of the shared predictors have statistically significant Wald tests (at the $\alpha = 0.05$ level). In addition, the parameter estimates for these variables have the same sign and order. AIC and BIC cannot be used to chose between these two models as fit.aic is optimal w.r.t AIC and fit.big is optimal w.r.t BIC.

```
summary(fit.aic)
```

```
##
## Call:
## glm(formula = HD ~ AGE + SEX + SBP + CHOL + FRW + CIG, family = binomial,
##     data = hd_data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7439  -0.7234  -0.5477  -0.3344   2.4607
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.013906   1.204650  -8.313  < 2e-16 ***
## AGE           0.070970   0.017749   3.999 6.37e-05 ***
## SEXMALE       0.841155   0.183481   4.584 4.55e-06 ***
## SBP           0.015800   0.003032   5.212 1.87e-07 ***
## CHOL          0.004878   0.001783   2.736  0.00622 **
## FRW           0.008490   0.004717   1.800  0.07190 .
## CIG           0.012740   0.007240   1.760  0.07845 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1053.82  on 999  degrees of freedom
## Residual deviance:  961.69  on 993  degrees of freedom
## AIC: 975.69
##
## Number of Fisher Scoring iterations: 4
```

```
summary(fit.bic)
```

```
##
## Call:
## glm(formula = HD ~ AGE + SEX + SBP + CHOL, family = binomial,
##     data = hd_data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6074  -0.7275  -0.5526  -0.3494   2.4449
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.936469   1.088535  -8.210  < 2e-16 ***
## AGE          0.065598   0.017410   3.768 0.000165 ***
## SEXMALE      0.905416   0.169629   5.338 9.42e-08 ***
```

```
## SBP          0.017078   0.002892   5.905 3.53e-09 ***
## CHOL         0.004848   0.001773   2.735 0.006241 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1053.82  on 999  degrees of freedom
## Residual deviance:  967.39  on 995  degrees of freedom
## AIC: 977.39
##
## Number of Fisher Scoring iterations: 4
```

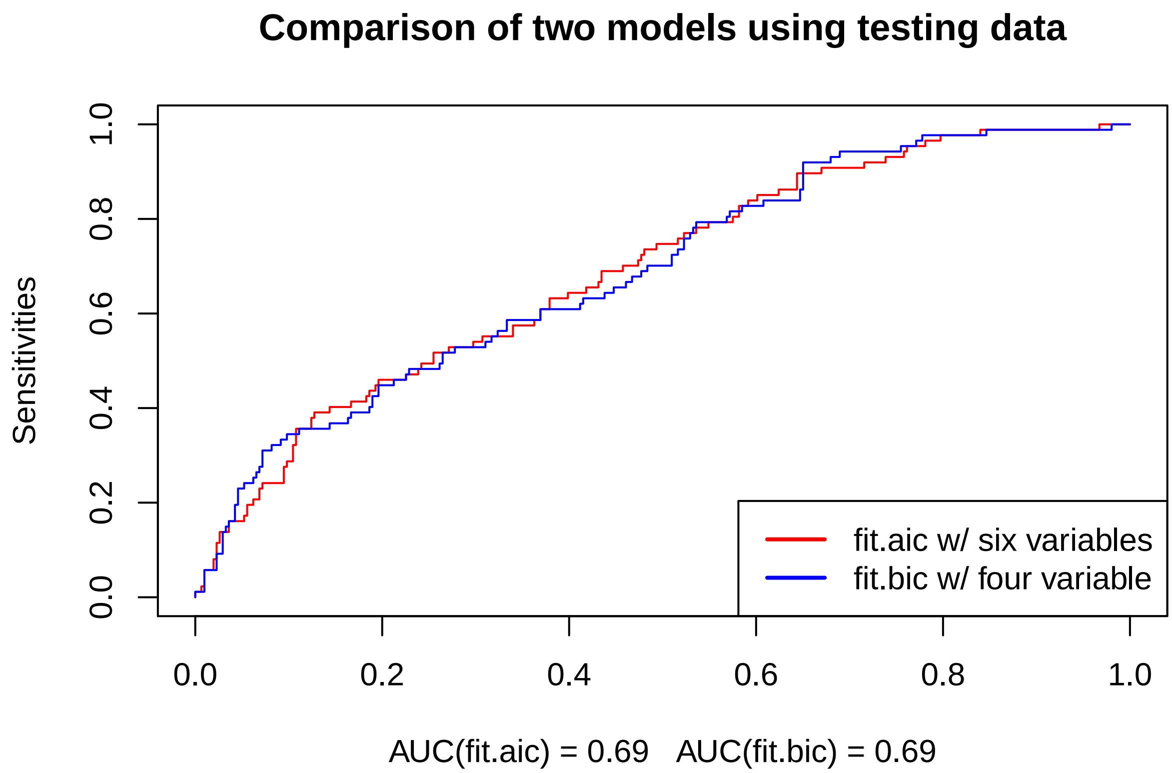
2. Overlay two ROC curves with the test dataset: One from `fit.bic`, the other from `fit.aic` from part A1. Based on the ROC curves, which one do you prefer?

Solution

Based upon the plot below I would prefer to use fit.bic, as it has the greater of the two AUC values (i.e. AUC(fit.bic)=0.69 > AUC(fit.aic)=0.68).

```
plot(1-fit.aic.roc$specificities, fit.aic.roc$sensitivities, col="red", pch=16,type="l",
     xlab=paste("AUC(fit.aic) =",
                round(pROC::auc(fit.aic.roc),2),
                "  AUC(fit.bic) =",
                round(pROC::auc(fit.bic.roc),2) ),
     ylab="Sensitivities")

points(1-fit.bic.roc$specificities, fit.bic.roc$sensitivities, col="blue",type="l", pch=16)
legend("bottomright", legend=c("fit.aic w/ six variables", "fit.bic w/ four variable"),
       lty=c(1,1), lwd=c(2,2), col=c("red", "blue"))
title("Comparison of two models using testing data")
```


Comparison of two models using testing data
Sensitivities
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
fit.aic w/ six variables
fit.bic w/ four variable
AUC(fit.aic) = 0.69   AUC(fit.bic) = 0.69