HW5 - 250620601

Ravin Lathigra

December 4, 2018

R Packages & Libraries

```
library(corrplot)    #Visualize Correlation between variables
library(kableExtra)  #Style tables
library(tidyverse)   #contains ggplot2,dplyr,tidyr, readr,purr,tibble,stringr,forcats
library(formatR)     #Improve readability of code
library(e1071)       #Functions for latent class analysis, Fourier transform ect.
library(VIM)         #Knn
library(caret)       #streamlined model development
library(ElemStatLearn)
library(nnet)
library(rattle.data)
library(pROC)
```

Question 1

Consider the following model:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$p = p(Y = 1|x_1, x_2)$$

Part A

Goal: Determine P(y=1|x1=1,x2=0.5)

$$p(Y = 1|x_1, x_2) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

```
## [1] 0.4218338
```

Given b0,b1,b2 = {-2.7399,3.0287,−1.2081} respectively, the probability that Y equals class 1 when x1 & x2 equal 1 and 0.5 respectively is: **0.4218338**

Part B

Goal: Test the following hypothesis:

null hypthesis:

$$H_o : \beta_2 = 0$$

and althernative hypthesis:

$$H_a : \beta_2 \neq 0$$

and

$$\alpha = 0.05$$

Hypothesis testing shows that given the z-value of B2, the probability of observing a value more exteme than the z-value is **less than** the signifcance level therefore we **reject** the null hypothesis.

Table 1: Question 1 B - Significance of B2

Parameter	z-value	pr(>|z|)	Action
B2	2.615	0.0089227	Reject Null

Part C

Goal: Test the following hypothesis:

null hypthesis:

$$H_o : \beta_1 = \beta_2 = 0$$

and althernative hypthesis:

$$H_a : H_0 is false$$

and

$$\alpha = 0.05$$

Steps:

Determine the D-Stat + Since there are only 2 predictors + intercept, we do not need to fit a reduced model, determine the deviance and take the difference between that and the full model. We can simply **compare the Null Deviance and Residual Deviance**.

$$D - stat = NullDeviance - ResidualDeviance$$

Determine the degrees of freedom + We are comparing the number of parameters in the Full and Reduced models. In our case, we are comparing a model with 3 predictors and a model with 1 i.e a difference of 2.

Determine Probability of Exceeding D-stat + When comparing full and reduced models, we know that this will follow a chi-squared distribution with k degrees of freedom, where k is the difference in predictors. From this we can determine the probability of observing a value D more extreme than our D-Stat.

Compare Probability to Significance Level + If the probability of exceeding the D-Stat is greater than the significance level of 0.05 we fail to reject the null hypothesis, otherwise we reject the null.

Table 2: Question 1 C - Significance of Model Predictors

D Stat	Degrees of Freedom	pr(>D)	Action
53.78	2	0	Reject Null

Hypothesis testing shows that given the D-Stat, the probability of observing a value more exteme than the D-Stat is **less than** the signifcance level therefore we **reject** the null hypothesis. At least one of B1 or B2 is significant for the model.

Question 2

Part A

```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 256  77
##          1  46  83
```

```
##
##                Accuracy : 0.7338
##                  95% CI : (0.691, 0.7735)
##     No Information Rate : 0.6537
##     P-Value [Acc > NIR] : 0.0001366
##
##                   Kappa : 0.3839
##  Mcnemar's Test P-Value : 0.0068303
##
##             Sensitivity : 0.8477
##             Specificity : 0.5188
##          Pos Pred Value : 0.7688
##          Neg Pred Value : 0.6434
##              Prevalence : 0.6537
##          Detection Rate : 0.5541
##    Detection Prevalence : 0.7208
##       Balanced Accuracy : 0.6832
##
##        'Positive' Class : 0
##
```

Table 3: Question 2a - Logistic Regression | Cutoff = 0.5

Measure	Result
Sensitivity	0.8476821
Precision	0.7687688
Accuracy	0.7337662
Specificity	0.5187500

Using the SAheart dataset, we can model `chd` using all of the predictors in the dataset.

The SAheart provides insight into South African Hearth Disease. Within the data we can identify `chd` (Corinary Hear Disease) as the reposnse variable that is dependent on the other predictors in the model. It is important to note that the reponse variable is a binary indivator whether an individual has corinary heart disease. While model *accuracy* is important, it is important that the implication of the prediction is considered i.e false positive and false negatives cannot be interpreted as equally incorrect.

The confusion matrix shoes how our model performed considering all predictors in the data. Table 3 seprates the accuracy, sensitivity, specificity and precision of the model, though the results are consistent with those in the confusion matrix output.

We gather that out model has a high sensitivity i.e if the diagnosis is negative (no chd) the model correctly assigns the diagnosis 84.77% of the time. oF all the negative classificaitons (no chd) made by the model the precision shows that it correctly classifies them 76.88% of the time. What we gather from the model is we better model none chd cases then we do chd, The specificity suggests that our prediciton only slightly outperforms a coin flip.

This is a good example of how accuracy can be misleading. While the model accuracy is high, it poorly models 1 class and more appropriately classfies the other.

Part B

Using backward selection with BIC, the best subset of predictors were `Tobacco`, `LDL`, `FamhistPresent`, `typea` and `age`.

Table 4 illustrate the coefficients of these predictors including the intercept.

Table 4: Question 2b - Logistic Regression backward selection | Cutoff = 0.5

	Coefficients
(Intercept)	-6.4464445
tobacco	0.0803753
ldl	0.1619916
famhistPresent	0.9081753
typea	0.0371152
age	0.0504604

Part C

```
## [1] 3.545546
```

Table 5: Question 2C - Significance of Model Predictors

D Stat	Degrees of Freedom
3.545546	4

Considering the full logistic regression and the BIC reduced model we can develop hypothesis to test.

null hypthesis:

$$H_o : \beta_{sbp} = \beta_{adiposity} = \beta_{obesity} = \beta_{alcohol} = 0$$

and althernative hypthesis:

$$H_a : \exists \;\; \beta_{sbp} = \beta_{adiposity} = \beta_{obesity} = \beta_{alcohol} \neq 0$$

and

$$\alpha = Significancelevel$$

Table 5 shows the D-Stat and Degrees of freedom used for the hypothesis testing.

Part D

Table 6: Question 2D - Significance of Model Predictors

D Stat	Degrees of Freedom	pr(>D)	Action
3.545546	4	0.4709869	Fail to Reject

Considering the a significance level of 5%, hypothesis testing suggests that we **fail to reject the null**.

Table 6 shows the probability of exceeding the D-statistic and the corresponding decision regarding the null.

Question 3

Part A

```
## Parsed with column specification:
## cols(
##   Age = col_integer(),
##   Gender = col_character(),
##   TB = col_double(),
##   DB = col_double(),
##   Alkphos = col_integer(),
```

```
##   Sgpt = col_integer(),
##   Sgot = col_integer(),
##   TP = col_double(),
##   ALB = col_double(),
##   A.G = col_double(),
##   Selector = col_integer()
## )
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 113  37
##          2  13  16
##
##                Accuracy : 0.7207
##                  95% CI : (0.6488, 0.785)
##     No Information Rate : 0.7039
##     P-Value [Acc > NIR] : 0.344648
##
##                   Kappa : 0.2287
##  Mcnemar's Test P-Value : 0.001143
##
##             Sensitivity : 0.8968
##             Specificity : 0.3019
##          Pos Pred Value : 0.7533
##          Neg Pred Value : 0.5517
##              Prevalence : 0.7039
##          Detection Rate : 0.6313
##    Detection Prevalence : 0.8380
##       Balanced Accuracy : 0.5994
##
##        'Positive' Class : 1
##
```

Table 7: Question 3a - Logistic Regression | Cutoff = 0.5

Measure	Result
Sensitivity	0.8968254
Precision	0.7533333
Accuracy	0.7206704
Specificity	0.3018868

The `ILPD` dataset provides insight into Indian Liver Patients. Within the data we can identify `Selector` - a class label used to divide into groups(liver patient or not)- as the reposnse variable that is dependent on the other predictors in the model. It is important to note that the reponse variable is a binary indicator. While model *accuracy* is important, it is important that the implication of the prediction is considered i.e false positive and false negatives cannot be interpreted as equally incorrect.

The confusion matrix shoes how our model performed considering all predictors in the data. Table 7 seprates the accuracy, sensitivity, specificity and precision of the model, though the results are consistent with those in the confusion matrix output.

We gather that out model has a high sensitivity i.e if `Selector` is 1 the model correctly assigns the diagnosis 89.68% of the time. oF all the Selecter = 1 classificaitons made by the model the precision shows that it

correctly classifies them 75.33% of the time. What we gather from the model is we better model selector group 1 cases then we do group 2. The specificity suggests that our prediciton performs significantly worse than a coin flip.

This is a good example of how accuracy can be misleading. While the model accuracy is moderate, it poorly models 1 class and more appropriately classfies the other.

Part B

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 125  53
##          2   1   0
##
##                Accuracy : 0.6983
##                  95% CI : (0.6254, 0.7646)
##     No Information Rate : 0.7039
##     P-Value [Acc > NIR] : 0.601
##
##                   Kappa : -0.0111
##  Mcnemar's Test P-Value : 3.915e-12
##
##             Sensitivity : 0.9921
##             Specificity : 0.0000
##          Pos Pred Value : 0.7022
##          Neg Pred Value : 0.0000
##              Prevalence : 0.7039
##          Detection Rate : 0.6983
##    Detection Prevalence : 0.9944
##       Balanced Accuracy : 0.4960
##
##        'Positive' Class : 1
##
```

Table 8: Question 3b - Logistic Regression | Cutoff = 0.8

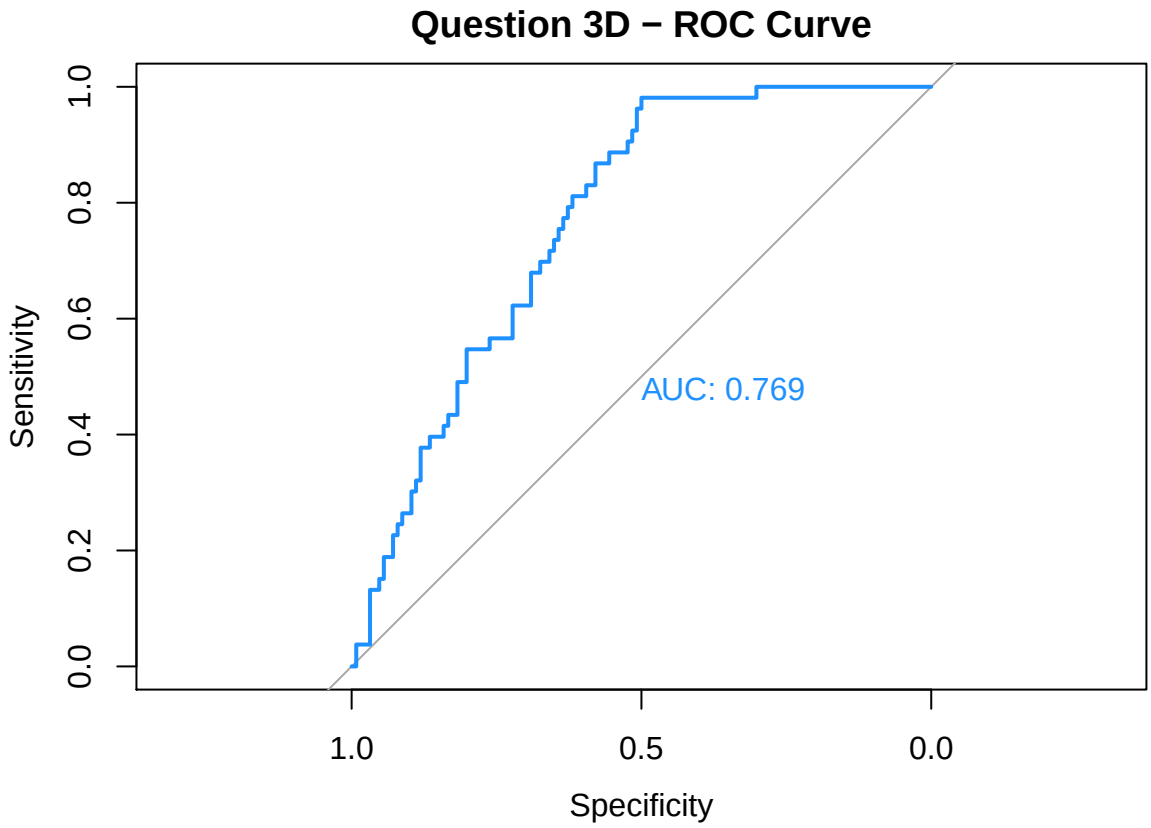
Measure	Result
Sensitivity	0.9920635
Precision	0.7022472
Accuracy	0.6983240
Specificity	0.0000000

Table 8 shows how our classification changes after changing the cutoff to 0.8. We notice that the sensitivity is near perfect but, we poorly represent group 2 cases. By increasing the cutoff, we put more restricition on what we consider to be "group 2". In only 1 case did the P(y=2|.) exceed 0.8.

Part C

As hinted at in Part B, to increase the sensitivity, we can increase the cutoff closer to 1. This increases the constraint to make a group 2 classification. If class 2 had been noted as the postiive class instead of class 1, the opposite would have been true, i.e decrease the cutoff.

Part D



AUC = 0.769

Part E

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

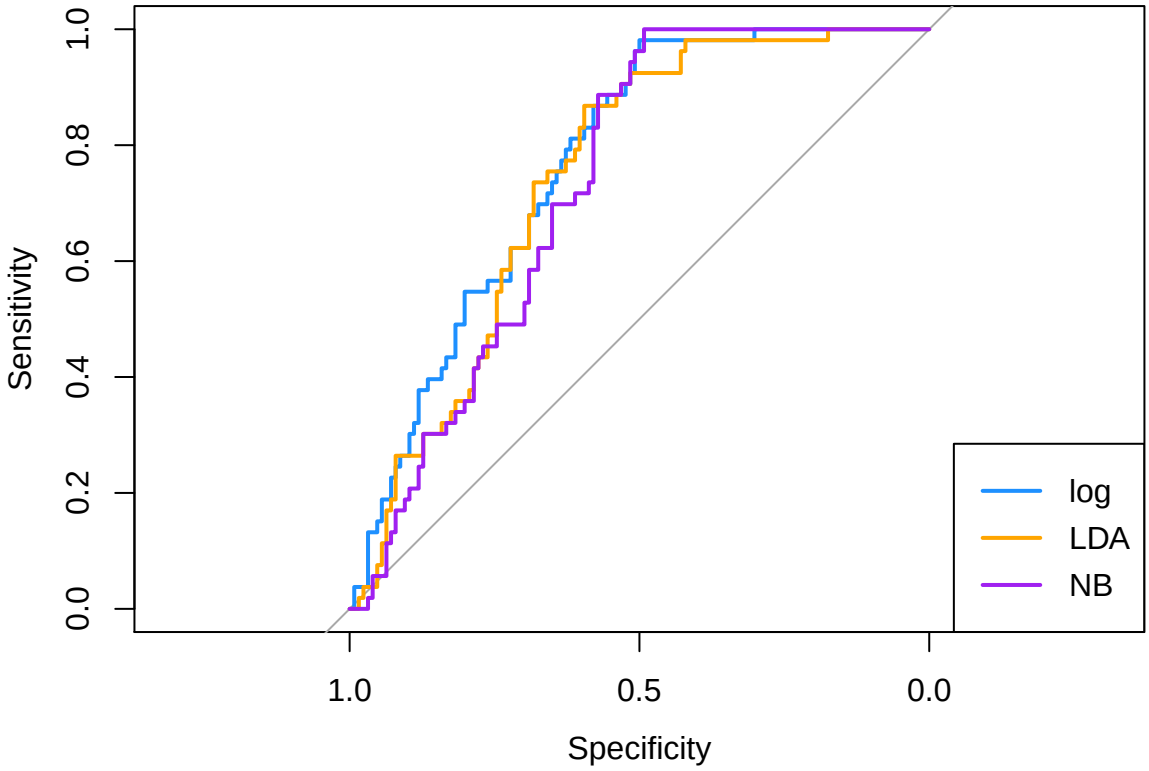


Table 9: Question 3E - AUC

	log	LDA	NB
AUC	0.7689428	0.7439353	0.7318059

Table 9 compares the AUC across the Logistic, Linear discriminant analysis and Naive Bayes models. We gather that the logistic regression has the largest AUC, followed by LDA and lastly NB.