Names: Gian David, Garrett Hedley, Jeffrey Medina, Jose Melendez, Alonzo Sule

You are going to use the Kaggle dataset framingham_heart_disease.csv obtained from https://www.kaggle.com/naveengowda16/logistic-regression-heart-disease-prediction

Throughout this homework your dependent variable is TenYearCHD

0) Read your framingham_heart_disease csv file, drop the rows with at least one missing value using df.dropna() function where df is the name of the dataframe you created when you read it in. For instance cleandata= df.dropna()

   Export this data as a CSV file for the questions below. For instance:
   np.savetxt("C:/Users/rm84/Desktop /deadmatrix.csv", cleandata, delimiter=",")
   Copy/paste here the script.

```
import pandas as pd
import numpy as np
heart = pd.read_csv("C:/Users/Garrett/Desktop/QMST 3339/framingham_heart_disease.csv", sep= ',')
cleandata= heart.dropna()
np.savetxt("C:/Users/Garrett/Desktop/CleanHeart.csv", cleandata, delimiter=",")
```

1) We will be logistic regression analysis using variables from the Kaggle data set. in doing so you will be selecting from all the candidate independent variables available in the kaggle data set. Use the AIC measure to determine which of the independent variables will be selected. All you need to do run logistic regression models obtain AIC. In the initial step and choose the independent variable that give you the minimum AIC. This is not an official approach but it allows us to start the model. Then you use the loop given to you in order to come up with a model that has the smallest possible AIC using the forward model building approach.

The code below are code snippets that are going to help you.

RCode

```
ccFraud<-
read.table(header=TRUE,file="C:/Users/rm84/Desktop/Teaching/datasets/creditca
rd.csv",sep=",")
attach(ccFraud)
```

####Useful for question a

```
names(ccFraud)
K=ncol(ccFraud)-1
AICF=matrix(nrow=K,ncol=1)
sAICF=matrix(nrow=K,ncol=1)
droppedc=matrix(nrow=K,ncol=1)
attach(ccFraud)
for(k in 1:K){
AICF[k]=summary(glm(Class~ccFraud[,k],family=binomial(link="logit")))$aic
}
ord=order(AICF)
```

```
####Useful for question b
sAICF[1]=AICF[which.min(AICF)]
space=ccFraud[,ord[1]]
for(k in 2:K){
 space=cbind(space,ccFraud[,ord[k]])
 lcol=dim(space)[2]
 sAICF[k]=summary(glm(Class~space,family=binomial(link="logit")))$aic
 if (sAICF[k]>=sAICF[k-1]){
  space=space[,-c(lcol)];droppedc[k]=ord[k]
 }
}
IVSpace=ccFraud[,c(setdiff(ord,droppedc))]
optima=glm(Class~. ,data=IVSpace,family=binomial(link="logit"))
##############Useful for question c and d
sum(optimapred== Class)/length(Class)
```

a) Having done the logistic regression in R , which is the variable with the lowest AIC, what is that AIC?

V14, 2921.433

b) Using the logistic regression in R, use the forward model building procedures for a logistic regression model.

```
Call:
glm(formula = Class ~ ., family = binomial(link = "logit"), data = IVSpace)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8558  -0.0299  -0.0193  -0.0113   4.4341

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.8334401  0.1577757 -55.987  < 2e-16 ***
V14         -0.5161062  0.0575517  -8.968  < 2e-16 ***
V12          0.1344795  0.0822039   1.636  0.10185
V10         -0.8550408  0.1000947  -8.542  < 2e-16 ***
V17         -0.0766781  0.0605847  -1.266  0.20564
V16         -0.2406154  0.1088251  -2.211  0.02703 *
V4           0.7921601  0.0627725  12.620  < 2e-16 ***
V3           0.0659176  0.0399187   1.651  0.09868 .
V18         -0.0052590  0.1211249  -0.043  0.96537
V2           0.1483501  0.0506451   2.929  0.00340 **
V5           0.1936908  0.0607624   3.188  0.00143 **
V6          -0.1372297  0.0706968  -1.941  0.05225 .
V21          0.3241444  0.0547692   5.918 3.25e-09 ***
V8          -0.1648094  0.0259453  -6.352 2.12e-10 ***
V20         -0.3587470  0.0643938  -5.571 2.53e-08 ***
V27         -0.6569666  0.1204117  -5.456 4.87e-08 ***
V28         -0.2545896  0.0889147  -2.863  0.00419 **
V13         -0.3467368  0.0792697  -4.374 1.22e-05 ***
Amount       0.0009986  0.0003090   3.231  0.00123 **
V22          0.5198676  0.1231052   4.223 2.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7242.5  on 284806  degrees of freedom
Residual deviance: 2246.3  on 284787  degrees of freedom
AIC: 2286.3

Number of Fisher Scoring iterations: 11
```

c) Obtain the accuracy of the model you built in b. Copy Paste your script and the accuracy value.

```
(sum((optimapred==1 & Class==1))+sum((optimapred==0 & Class==0)))/length(Class)
```

0.9991924

d) Obtain the sensitivity of the model you built in b. Copy Paste your script and the sensitivity value.

```
optimapred=(predict(optima,type="response")>0.5)*1
sum((optimapred==1 & Class==1))/sum(Class==1)
```

0.6178862