Logistic Regression Analysis on Heart Health Data

The Issue:

The given dataset related to heart health, which includes 18 factors including the "delay days" variable, measured in days until a person seeks medical treatment. Our objective is to create a logistic regression model to predict whether a person seeks medical treatment in three ways. The first method involves categorizing those who seek treatment in 2 days or less as "1" and those who seek treatment after 2 days as "0". The second method involves calculating the average delay time for a group of people and labeling those who seek treatment in less than the average time as "1" and those who seek treatment after the average time as "0". The third method involves labeling those who seek treatment in less than 1 day as "1" and those who seek treatment after 1 day as "0".

Findings:

We are developing a logistic regression model using 18 variables, including "delay days," to determine the significance of these variables in predicting delay days. We can evaluate the statistical significance of each variable using their corresponding p-values and z-values. Our results show that only "cough" has a statistically significant association with delay days, with a p-value of 0.000505 and edema 0.007666. Variables with p-values less than 0.05 are considered statistically significant and indicate a non-chance association with the dependent variable. Therefore, we can conclude that there is a significant association between cough, edema and delay days, while the remaining variables (ID, age, gender, ethnicity, marital status, living situation, education, palpitations, orthopnea, chest pain, nausea, fatigue, dyspnea, PND, tight shoes, weight gain, and DOE) are not significant predictors of delay days. We also generated a ROC curve to evaluate the performance of our model, which resulted in an ROC-AUC value of 0.581, suggesting that our model's performance is satisfactory.

Discussions:

When interpreting the results of the logistic model with 18 variables, it is important to consider the corresponding p-values. A p-value greater than 0.05 suggests that the variable has no significant impact, while a p-value less than 0.05 suggests that the variable is significant to some extent. The coefficient values provide insights into the direction and magnitude of the relationship between the predictor and dependent variables. A positive coefficient indicates that an increase in the predictor variable is associated with an increase in the dependent variable, and a negative coefficient suggests the opposite. The ROC curve and AUC value are useful in evaluating the model's performance. Based on these metrics, we can determine if the model's performance is satisfactory or poor.

Appendix A Method:

To build our logistic model, we used the R programming language and installed relevant packages such as pROC and caTools. We imported the dataset and separated the dependent variable, delay days, from the other variables. We created a new dataset with the remaining variables and split it into training and testing data sets. Using the training data, we constructed the generalized logistic model and applied it to generate predictions with the testing data. We analyzed the model predictions to determine the significant variables. We evaluated the effectiveness of our model by creating a ROC curve using the generalized logistic model and calculating the AUC. We also obtained the confusion matrix to calculate the accuracy.

Appendix B: Result

Results of generalized logistic model for case 1

```
> library(readxl)
> library(pROC)
> file <- "C:\\Users\\sasik\\OneDrive\\Desktop\\heart-health-data.xls"
> data <- read_excel(file, sheet = 1)
> data$delay<-ifelse(data$delaydays<2,1,0)
> summary(data)
       ID             Age             Gender         Ethnicity         Marital        
Livewith
 Min.   :  1.0   Min.   :41.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Mi
n.   :1.000
```

```
 1st Qu.: 52.0   1st Qu.:67.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1s
t Qu.:1.000
 Median :103.5   Median :77.00   Median :1.000   Median :1.000   Median :2.000   Me
dian :2.000
 Mean   :127.9   Mean   :74.25   Mean   :1.448   Mean   :1.126   Mean   :1.665   Me
an   :1.663
 3rd Qu.:189.8   3rd Qu.:84.00   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:2.000   3r
d Qu.:2.000
 Max.   :512.0   Max.   :96.00   Max.   :2.000   Max.   :5.000   Max.   :3.000   Ma
x.   :2.000
                                                                                 NA
's   :2
   Education      palpitations      orthopnea       chestpain         nausea
 Min.   :1.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.0000
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0000
 Median :2.000   Median :0.0000   Median :2.000   Median :0.000   Median :0.0000
 Mean   :2.118   Mean   :0.6478   Mean   :1.638   Mean   :0.564   Mean   :0.5025
 3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:1.0000
 Max.   :6.000   Max.   :3.0000   Max.   :3.000   Max.   :3.000   Max.   :3.0000

     cough          fatigue         dyspnea          edema             PND
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:0.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:0.000
 Median :1.000   Median :2.000   Median :2.000   Median :1.000   Median :1.000
 Mean   :1.081   Mean   :1.951   Mean   :1.995   Mean   :1.227   Mean   :1.286
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :3.000

   tightshoes       weightgain          DOE          delaydays           delay
 Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :  0.000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:  0.500   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :2.000   Median :  2.000   Median :0.0000
 Mean   :0.8399   Mean   :0.8892   Mean   :1.739   Mean   :  5.726   Mean   :0.4367
 3rd Qu.:2.0000   3rd Qu.:2.0000   3rd Qu.:2.750   3rd Qu.:  7.000   3rd Qu.:1.0000
 Max.   :3.0000   Max.   :3.0000   Max.   :3.000   Max.   :225.000   Max.   :1.0000
                                                   NA's   :3         NA's   :3
>
> data1 <- subset(data,select = -delaydays)
>
> #Spliting the data
>
> div<-sample(2,nrow(data1),replace=T,prob=c(0.7,0.3))
> training<-data1[div==1,]
> testing<-data1[div==2,]
>
>
>
> lm<-glm(delay~.,data=training,family='binomial')
> summary(lm)

Call:
glm(formula = delay ~ ., family = "binomial", data = training)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.4254  -0.9968  -0.6028   1.0498   1.9553  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -0.9134731  1.5437641  -0.592 0.554039    
ID           -0.0009859  0.0014254  -0.692 0.489161    
Age           0.0157609  0.0122649   1.285 0.198779    
Gender        0.2951178  0.2776102   1.063 0.287752    
Ethnicity    -0.1772361  0.2467537  -0.718 0.472590    
Marital       0.4340413  0.2316556   1.874 0.060979 .  
Livewith      0.0925823  0.3369537   0.275 0.783499    
Education    -0.0361904  0.0987332  -0.367 0.713957    
palpitations  0.1373554  0.1558103   0.882 0.378017    
orthopnea    -0.1218716  0.1520376  -0.802 0.422791    
chestpain     0.3399483  0.1651191   2.059 0.039513 *  
nausea       -0.0181303  0.1686904  -0.107 0.914411    
cough        -0.5183568  0.1490328  -3.478 0.000505 ***
fatigue      -0.2322050  0.1743744  -1.332 0.182977    
dyspnea       0.1130933  0.1665321   0.679 0.497069    
edema        -0.4238133  0.1589450  -2.666 0.007666 ** 
PND          -0.0558181  0.1471694  -0.379 0.704482    
tightshoes    0.2206363  0.1677055   1.316 0.188302    
```

```
weightgain     0.1914854  0.1473069   1.300 0.193633
DOE           -0.2504334  0.1566355  -1.599 0.109859
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 377.73  on 274  degrees of freedom
Residual deviance: 333.15  on 255  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 373.15

Number of Fisher Scoring iterations: 3

>
> pre<-predict(lm, testing, type='response')
>
> #ROC curve
> ROC <- roc(testing$delay,pre)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> plot(ROC , print.auc= TRUE)
>
> #Confusion Matrix
> pre1<-ifelse(pre>0.5,1,0)
>
> table(pre1)
pre1
 0  1
75 53
> tab<-table(Prediction=pre1,Actual=testing$delay)
>
> #Accuracy,Misclassification error
> Accuracy<-sum(diag(tab))/sum(tab)
> error<-1-Accuracy
> error
[1] 0.4047619
```

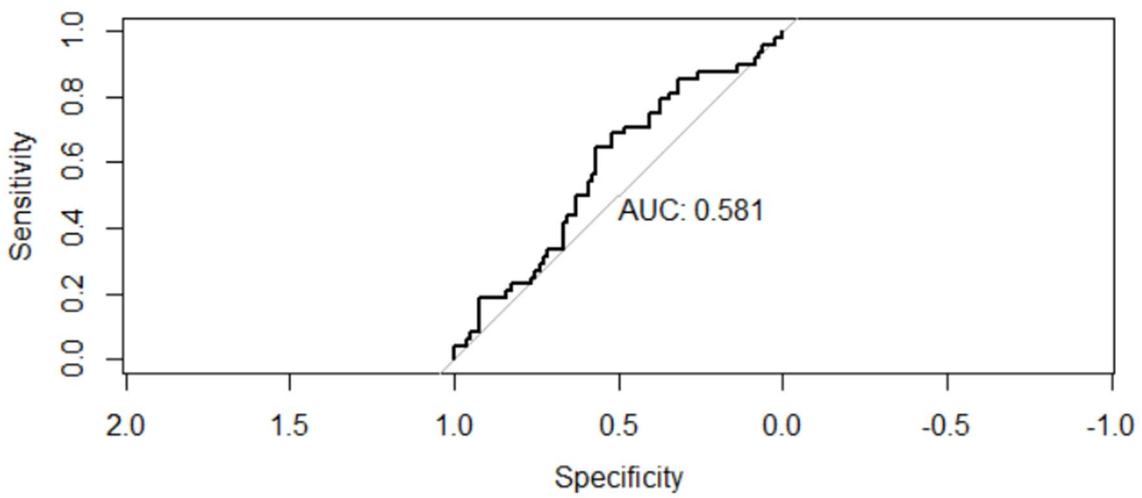

Fig1 : ROC curve and AUC for case 1.

Results of generalized logistic model for case 2

```
> file <-"C:\\Users\\sasik\\OneDrive\\Desktop\\heart-health-data.xls"
```

```
> data <- read_excel(file, sheet = 1)
> mean_d
[1] 5.725779
> data$delay<-ifelse(data$delaydays<mean_d,1,0)
>
> data1<-subset(data,select = -delaydays)
>
> #Spliting the data
> div<-sample(2,nrow(data1),replace=T,prob=c(0.7,0.3))
> training<-data1[div==1,]
> testing<-data1[div==2,]
>
> #logistic model
> lm<-glm(delay~.,data=training,family='binomial')
> summary(lm)

Call:
glm(formula = delay ~ ., family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9902  -1.1513   0.6622   0.8230   1.4104

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.439567   1.555640   1.568   0.1168
ID           -0.002366   0.001423  -1.662   0.0965 .
Age          -0.002292   0.011759  -0.195   0.8455
Gender        0.059303   0.284840   0.208   0.8351
Ethnicity    -0.163910   0.217080  -0.755   0.4502
Marital       0.030560   0.241537   0.127   0.8993
Livewith     -0.047014   0.352192  -0.133   0.8938
Education     0.006434   0.107460   0.060   0.9523
palpitations -0.149900   0.156099  -0.960   0.3369
orthopnea    -0.005254   0.150863  -0.035   0.9722
chestpain     0.084100   0.156627   0.537   0.5913
nausea       -0.388581   0.172856  -2.248   0.0246 *
cough        -0.033916   0.142562  -0.238   0.8120
fatigue      -0.013613   0.176807  -0.077   0.9386
dyspnea      -0.032985   0.174692  -0.189   0.8502
edema        -0.376465   0.151807  -2.480   0.0131 *
PND          -0.168877   0.137572  -1.228   0.2196
tightshoes    0.237411   0.164488   1.443   0.1489
weightgain    0.089100   0.142861   0.624   0.5328
DOE          -0.053832   0.162044  -0.332   0.7397
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 346.58  on 288  degrees of freedom
Residual deviance: 326.88  on 269  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 366.88

Number of Fisher Scoring iterations: 4


>
> #Prediction
> pre<-predict(lm, testing, type='response')
> pre
        1         2         3         4         5         6         7         8
9
0.8408764 0.5141380 0.8736205 0.8467192 0.8467192 0.9064892 0.8384055 0.7238536 0.7
706020
       10        11        12        13        14        15        16        17
18
0.8287406 0.8231606 0.8407530 0.8819686 0.8766920 0.9037667 0.3767912 0.8737457 0.7
975919
       19        20        21        22        23        24        25        26
27
0.7494097 0.8130017 0.7799863 0.7884709 0.8061686 0.7780027 0.7491107 0.8080437 0.6
425924
       28        29        30        31        32        33        34        35
36
0.8111050 0.5743026 0.6826423 0.7325440 0.7581208 0.7661918 0.8142332 0.8466834 0.3
866615
```

```
       37         38         39         40         41         42         43         44 
45 
0.7919466 0.8249310 0.7887095 0.6876546 0.7454760 0.7310816 0.6122362 0.5918874 0.6
055657 
       46         47         48         49         50         51         52         53 
54 
       NA 0.8810881 0.6295598 0.8636308 0.8311080 0.8308963 0.8799356 0.8286788 0.4
692084 
       55         56         57         58         59         60         61         62 
63 
0.5956719 0.6843973 0.7545373 0.6128855 0.5470719 0.6885544 0.6783366 0.7872351 0.7
555715 
       64         65         66         67         68         69         70         71 
72 
0.6917214 0.6504256 0.6085514 0.5931528 0.6647217 0.8646614 0.6013581 0.5971433 0.7
933604 
       73         74         75         76         77         78         79         80 
81 
0.7454263 0.6463831 0.6737098 0.7637558 0.6941591 0.7071066 0.7675496 0.5278110 0.7
684027 
       82         83         84         85         86         87         88         89 
90 
0.7372016 0.7065059 0.4858055 0.8645326 0.7047265 0.6772332 0.6327062 0.6124889 0.6
002613 
       91         92         93         94         95         96         97         98 
99 
0.5338180 0.7743471 0.7006888 0.6805711 0.8117873 0.5025104 0.4695909 0.7833172 0.6
396541 
      100        101        102        103        104        105        106        107 
108 
0.7797908 0.7878563 0.6109720 0.7614416 0.5384182 0.6458852 0.6021039 0.5676860 0.8
208245 
      109        110        111        112        113        114 
0.5530519 0.4739046 0.6348737 0.5519041 0.4773750 0.6312300 
> 
> #ROC curve
> ROC <- roc(testing$delay,pre)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> plot(ROC , print.auc= TRUE)
> 
> #Confusion Matrix
> pre1<-ifelse(pre>0.5,1,0)
> 
> table(pre1)
pre1
  0   1 
  7 106 
> tab<-table(Prediction=pre1,Actual=testing$delay)
> 
> 
> #Accuracy,Misclassification error
> Accuracy<-sum(diag(tab))/sum(tab)
> M_error<-1-Accuracy
> M_error
[1] 0.3125
```

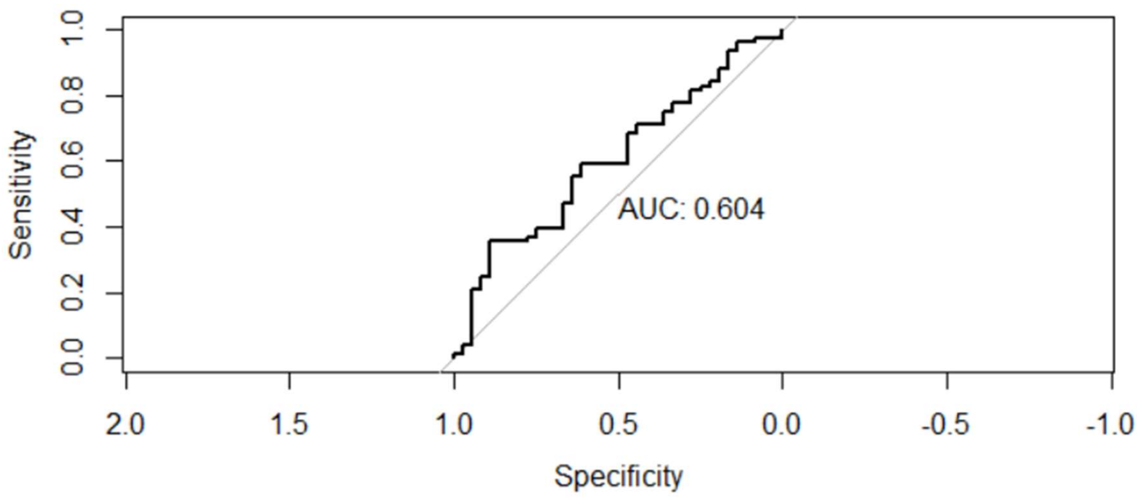

Fig2: ROC curve and AUC for case 2

Results of generalized logistic model for case 3

```
> file <-"C:\\Users\\sasik\\OneDrive\\Desktop\\heart-health-data.xls"
> data <- read_excel(file, sheet = 1)
>
>
> data$delay<-ifelse(data$delaydays<1,1,0)
>
>
> #subset of original dataset by removing delaydays column
> data1 <- subset(data,select = -delaydays)
>
>
> #Spliting the data
> div<-sample(2,nrow(data1),replace=T,prob=c(0.7,0.3))
> training<-data1[div==1,]
> testing<-data1[div==2,]
>
>
> #logistic model
> lm<-glm(delay~.,data=training,family='binomial')
> summary(lm)

Call:
glm(formula = delay ~ ., family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4299  -0.8535  -0.6787   1.2178   2.0456

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2255233  1.6063376   0.763   0.4455
ID            0.0001782  0.0014432   0.123   0.9017
Age           0.0088701  0.0122514   0.724   0.4691
Gender       -0.0830137  0.2862066  -0.290   0.7718
Ethnicity    -0.2676876  0.2815853  -0.951   0.3418
Marital      -0.0146799  0.2447695  -0.060   0.9522
Livewith     -0.7626177  0.3503349  -2.177   0.0295 *
Education    -0.0265897  0.1018326  -0.261   0.7940
palpitations -0.0498858  0.1634236  -0.305   0.7602
orthopnea    -0.1417594  0.1531118  -0.926   0.3545
chestpain    -0.0338475  0.1709948  -0.198   0.8431
```

```
nausea        0.1815915  0.1764002   1.029   0.3033
cough        -0.2698629  0.1532091  -1.761   0.0782 .
fatigue       0.0132373  0.2010081   0.066   0.9475
dyspnea      -0.0163399  0.1722430  -0.095   0.9244
edema        -0.2867574  0.1723288  -1.664   0.0961 .
PND           0.0010417  0.1490358   0.007   0.9944
tightshoes   -0.0202963  0.1795442  -0.113   0.9100
weightgain    0.1585074  0.1450611   1.093   0.2745
DOE          -0.1460601  0.1577641  -0.926   0.3545
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 337.12  on 270  degrees of freedom
Residual deviance: 314.81  on 251  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 354.81

Number of Fisher Scoring iterations: 4

>
> #Prediction
> pre<-predict(lm, testing, type='response')
> pre
        1         2         3         4         5         6         7         8 
9
0.4818997 0.4529078 0.1735221 0.6824983 0.5116211 0.6159579 0.6159579 0.4099323 0.2
033604
       10        11        12        13        14        15        16        17 
18
0.2055133 0.2573223 0.2256162 0.4264937 0.3822858 0.2586502 0.1470329 0.6162410 0.7
756967
       19        20        21        22        23        24        25        26 
27
0.3990624 0.3990624 0.4581612 0.3968360 0.2092545 0.4309142 0.2233529 0.4872629 0.6
692507
       28        29        30        31        32        33        34        35 
36
0.1339520 0.2242483 0.2242793 0.3379722 0.4181390 0.3081403 0.3261531 0.2065905 0.4
271065
       37        38        39        40        41        42        43        44 
45
0.2638761 0.6444494 0.2101727 0.1145979 0.1810783 0.2847326 0.1392293 0.3054740 0.1
392507
       46        47        48        49        50        51        52        53 
54
0.1841986 0.2049819 0.1461525 0.1475365 0.2945313 0.1287133 0.3177139 0.3275188 0.2
414470
       55        56        57        58        59        60        61        62 
63
0.1742215 0.1605154 0.2601188 0.4729814 0.5135962 0.1252781 0.4496249 0.3554650 0.3
555059
       64        65        66        67        68        69        70        71 
72
0.3527325 0.3301402 0.3155709 0.2331018 0.1575646 0.2221425 0.2132824 0.2069375 0.1
716198
       73        74        75        76        77        78        79        80 
81
0.1765113 0.2715978 0.2780415 0.3109406 0.2934760 0.3038894 0.5075611 0.1586880 0.2
932055
       82        83        84        85        86        87        88        89 
90
0.2627392 0.2469783 0.3278462 0.6141103 0.2519227 0.3925725 0.3548536 0.4217869 0.3
137012
       91        92        93        94        95        96        97        98 
99
0.7340313 0.2658716 0.5388209 0.3853106 0.3227088 0.3125014 0.2254272 0.1484410 0.4
785078
      100       101       102       103       104       105       106       107 
108
0.1205976 0.2531022 0.2995725 0.2190199 0.5403961 0.3080409 0.2278974 0.3605842 0.1
386445
      109       110       111       112       113       114       115       116 
117
0.5084515 0.2475581 0.1318140 0.2558582 0.5155247 0.3493073 0.4073580 0.3276295 0.2
223984
```

```
      118       119       120       121       122       123       124       125
126
0.2387496 0.2178770 0.1509393 0.1820763 0.3864026 0.3014749 0.3396654 0.4117844 0.3
508576
      127       128       129       130       131
0.3247837 0.5261205 0.3367690 0.5041361 0.4548994
>
> #ROC curve
> ROC <- roc(testing$delay,pre)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> plot(ROC , print.auc= TRUE)
>
> #Confusion Matrix
> pre1<-ifelse(pre>0.5,1,0)
> pre1
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21
22  23  24
  0   0   0   1   1   1   1   0   0   0   0   0   0   0   0   0   1   1   0   0   0
0   0   0
 25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45
46  47  48
  0   0   1   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0
0   0   0
 49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69
70  71  72
  0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
0   0   0
 73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93
94  95  96
  0   0   0   0   0   0   1   0   0   0   0   0   1   0   0   0   0   0   1   0   1
0   0   0
 97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117
118 119 120
  0   0   0   0   0   0   0   1   0   0   0   0   1   0   0   0   1   0   0   0   0
0   0   0
121 122 123 124 125 126 127 128 129 130 131
  0   0   0   0   0   0   0   1   0   1   0
> table(pre1)
pre1
  0   1
113  18
> tab<-table(Prediction=pre1,Actual=testing$delay)
> tab
          Actual
Prediction  0  1
         0 82 31
         1 11  6
>
> #Accuracy,Misclassification error
> Accuracy<-sum(diag(tab))/sum(tab)
> M_error<-1-Accuracy
> M_error
[1] 0.3230769
```

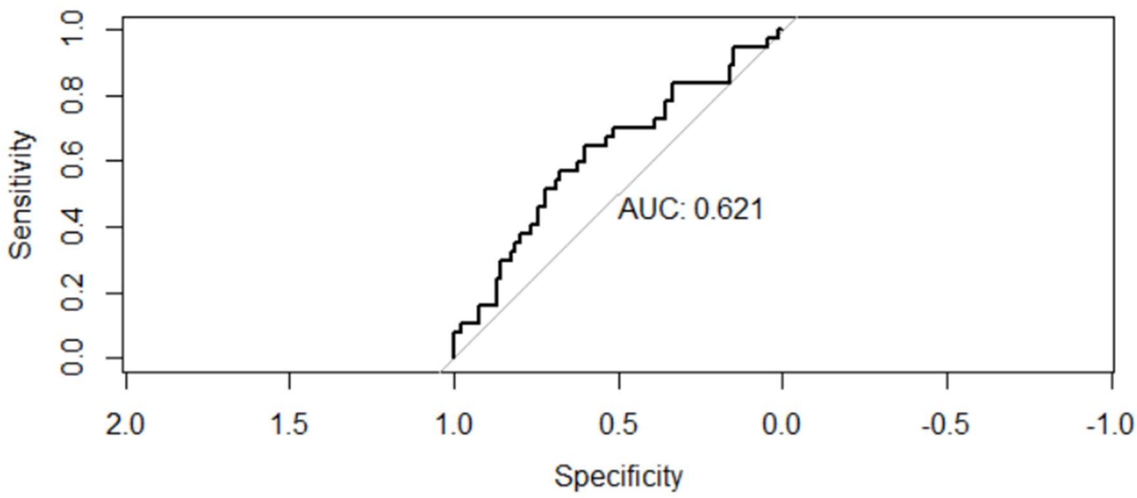

Fig3 : ROC curve and AUC for case 3

Appendix C Code:

Case1 :

```
library(readxl)
library(pROC)
file <- "C:\\Users\\sasik\\OneDrive\\Desktop\\heart-health-data.xls"
data <- read_excel(file, sheet = 1)
data$delay<-ifelse(data$delaydays<2,1,0)
summary(data)

data1 <- subset(data,select = -delaydays)

#Spliting the data
div<-sample(2,nrow(data1),replace=T,prob=c(0.7,0.3))
training<-data1[div==1,]
testing<-data1[div==2,]

lm<-glm(delay~.,data=training,family='binomial')
summary(lm)
```

```
pre<-predict(lm, testing, type='response')

#ROC curve
ROC <- roc(testing$delay,pre)
plot(ROC , print.auc= TRUE)

#Confusion Matrix
pre1<-ifelse(pre>0.5,1,0)

table(pre1)
tab<-table(Prediction=pre1,Actual=testing$delay)

#Accuracy,Misclassification error
Accuracy<-sum(diag(tab))/sum(tab)
error<-1-Accuracy
error
```

Case2 :

```
file <-"C:\\Users\\sasik\\OneDrive\\Desktop\\heart-health-data.xls"
data <- read_excel(file, sheet = 1)
data
str(data)

#mean for delaydays
mean_d<-mean(data$delaydays,na.rm=TRUE)
mean_d
data$delay<-ifelse(data$delaydays<mean_d,1,0)
```

```
data1<-subset(data,select = -delaydays)

#Spliting the data
div<-sample(2,nrow(data1),replace=T,prob=c(0.7,0.3))
training<-data1[div==1,]
testing<-data1[div==2,]

#logistic model
lm<-glm(delay~.,data=training,family='binomial')
summary(lm)

#Prediction
pre<-predict(lm, testing, type='response')
pre

#ROC curve
ROC <- roc(testing$delay,pre)
plot(ROC , print.auc= TRUE)

#Confusion Matrix
pre1<-ifelse(pre>0.5,1,0)

table(pre1)
tab<-table(Prediction=pre1,Actual=testing$delay)

#Accuracy,Misclassification error
Accuracy<-sum(diag(tab))/sum(tab)
M_error<-1-Accuracy
M_error
```

Case3 :

```
file <-"C:\\Users\\sasik\\OneDrive\\Desktop\\heart-health-data.xls"
data <- read_excel(file, sheet = 1)

data$delay<-ifelse(data$delaydays<1,1,0)

#subset of original dataset by removing delaydays column
data1 <- subset(data,select = -delaydays)

#Spliting the data
div<-sample(2,nrow(data1),replace=T,prob=c(0.7,0.3))
training<-data1[div==1,]
testing<-data1[div==2,]

#logistic model
lm<-glm(delay~.,data=training,family='binomial')
summary(lm)

#Prediction
pre<-predict(lm, testing, type='response')
pre

#ROC curve
```

```
ROC <- roc(testing$delay,pre)
plot(ROC , print.auc= TRUE)

#Confusion Matrix
pre1<-ifelse(pre>0.5,1,0)
pre1
table(pre1)
tab<-table(Prediction=pre1,Actual=testing$delay)
tab

#Accuracy,Misclassification error
Accuracy<-sum(diag(tab))/sum(tab)
M_error<-1-Accuracy
M_error
```