IS507 - Data, Statistical Models, and Information - Assignment 8

Problem 1: The Excel spreadsheet Alzheimer.csv contains one sheet named Alzheimer, which is data attempting to explain whether a patient has Alzheimer's Disease. These are data from a sample of 336 employees and consists of 9 variables for each patient. These are:

1) Dementia-Outcome variable-patient diagnosis
2) Gender-Female=0 and Male=1
3) Age-Age of patient (in years)
4) Education-Years of Education
5) SES-Socioeconomic Status 1=Low and 5=High
6) MMSE-Mini mental state examination score
7) CDR-Clinical Dementia Rating
8) eTIV-estimated total intracranial volume
9) nWBV-Normalize whole brain volume
10) ASF-Atlas Scaling Factor

Develop a **Logistic Regression model** to classify the Dementia event from the other variables.

Dataset Cleaning:

```r
# Load library readr to read the dataset
library(readr)
library(plyr)
library(dplyr)
library(tidyr)
library(stringr)

# Set working directory
setwd("~/Desktop/IS507 - Data, Statistical Models and Information/Assignments/Assignment_8")

# Import the dataset using read_csv()
alzheimer <- read.csv("alzheimer-1.csv", header=TRUE)
alzheimer <- alzheimer

# Shape of dataset
dim(alzheimer)

# Count of missing values
sum(is.na(alzheimer))

# Delete rows with missing values
new_alzheimer <- na.omit(alzheimer)

# Shape of new dataset after listwise deletion
dim(new_alzheimer)

# Renaming columns
names(new_alzheimer)[4] <- "Years_of_Education"
names(new_alzheimer)[5] <- "Socioeconomic_Status"
names(new_alzheimer)[6] <- "Mini_Mental_State_Examination_Score"
names(new_alzheimer)[7] <- "Clinical_Dementia_Rating"
names(new_alzheimer)[8] <- "Estimated_Total_Intracranial_Volume"
names(new_alzheimer)[9] <- "Normalize_Whole_Brain_Volume"
names(new_alzheimer)[10] <- "Atlas_Scaling_Factor"

str(new_alzheimer)
```

a) Create a logistic regression model and explain the significant odds ratios in terms of Dementia.

Solution:

- In preprocessing, I have converted the output Dementia variable into a factor variable with values as Alzheimer and No Alzheimer.
- The Age variable is not normal and hence we have divided into 3 tertiles with categories of Age Between 60 and 73, Age Between 74 and 86 and Age Between 87 and 98.
- I have converted the Gender binary variable into categorical with 1 being Male and 0 being Female for understanding the chances of Dementia with gender.

We create a Logistic Regression model with all the input variables except CDR because when we use it, we get odds ratio for it as a very large number which is something absurd.

We can see from the output that:

- Variable **Gender_Category with gender as Male** has p-value < 0.001 and so it is significant factor in determining if a patient has Alzheimer or not. The odds ratio for it is 0.15 < 1 and hence it is a preventative factor. This means that it is not a likely event that a male person may have Alzheimer.
- Variable **Age_Group with value as 'Age Between 73 and 85'** has p-value = 0.12 and so it is not a significant factor in determining if a patient has Alzheimer or not.
- Variable **Age_Group with value as 'Age Between 87 and 98'** has p-value < 0.001 and so it is significant factor in determining if a patient has Dementia or not. The odds ratio for it is 6.98 > 1 and hence it is a risk factor. This means that it is highly likely event that a person with age between 87 and 98 may have Alzheimer.
- Variable **Years_of_Education** has p-value 0.004 and so it is significant factor in determining if a patient has Dementia or not. The odds ratio for it is 1.19 > 1 and hence it is a risk factor. This means that it is highly likely event that a person having high Education may have Alzheimer.
- Variable **Socioeconomic_Status** has p-value 0.018 and so it is significant factor in determining if a patient has Dementia or not. The odds ratio for it is 1.41 > 1 and hence it is a risk factor. This means that it is highly likely event that a person having any kind of Socio Economic Status may have Alzheimer.
- Variable **Mini_Mental_State_Examination_Score** has p-value < 0.001 and so it is significant factor in determining if a patient has Dementia or not. The odds ratio for it is 3.11 > 1 and hence it is a risk factor. This means that it is highly likely event that a person having a Mini mental state examination score may have Alzheimer.
- Variable **Estimated_Total_Intracranial_Volume** has p-value = 0.9 and so it is not significant factor in determining if a patient has Dementia or not.
- Variable **Atlas_Scaling_Factor** has p-value 0.6 and so it is not significant factor in determining if a patient has Dementia or not.

```
> # Converting the Dementia variable to Factor
> new_alzheimer$Dementia <- as.factor(new_alzheimer$Dementia)
> table(new_alzheimer$Dementia)

   Alzheimer No Alzheimer
         381          570
>
> new_alzheimer$Age_Group[(new_alzheimer$Age >= 60) & (new_alzheimer$Age <= 72)] <- "Age Between 60 and 72"
> new_alzheimer$Age_Group[(new_alzheimer$Age > 72) & (new_alzheimer$Age <= 85)] <- "Age Between 73 and 85"
> new_alzheimer$Age_Group[(new_alzheimer$Age > 85) & (new_alzheimer$Age <= 98)] <- "Age Between 86 and 98"
> table(new_alzheimer$Age_Group)

Age Between 60 and 72 Age Between 73 and 85 Age Between 86 and 98
                  294                   525                   132
>
> new_alzheimer$Gender_Category[(new_alzheimer$Gender == 1)] <- "Male"
> new_alzheimer$Gender_Category[(new_alzheimer$Gender == 0)] <- "Female"
> table(new_alzheimer$Gender_Category)

Female   Male
   540    411
```

```
>
> # Dropping the CDR and nWBV columns as the odds-ratio is very absurd for them
> log_reg <- glm(
+   Dementia ~ Gender_Category + Age_Group + Years_of_Education +
+              Socioeconomic_Status + Mini_Mental_State_Examination_Score +
+              Estimated_Total_Intracranial_Volume + Atlas_Scaling_Factor,
+   family = "binomial",
+   data = new_alzheimer
+ )
> summary(log_reg)

Call:
glm(formula = Dementia ~ Gender_Category + Age_Group + Years_of_Education +
    Socioeconomic_Status + Mini_Mental_State_Examination_Score +
    Estimated_Total_Intracranial_Volume + Atlas_Scaling_Factor,
    family = "binomial", data = new_alzheimer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6352  -0.1228   0.2564   0.5002   2.6099

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -3.171e+01  1.146e+01  -2.767  0.00566 **
Gender_CategoryMale                 -1.891e+00  2.815e-01  -6.718 1.84e-11 ***
Age_GroupAge Between 73 and 85       3.728e-01  2.371e-01   1.572  0.11586
Age_GroupAge Between 86 and 98       1.942e+00  4.294e-01   4.524 6.07e-06 ***
Years_of_Education                   1.767e-01  6.135e-02   2.880  0.00398 **
Socioeconomic_Status                 3.445e-01  1.461e-01   2.357  0.01842 *
Mini_Mental_State_Examination_Score  1.134e+00  8.943e-02  12.685  < 2e-16 ***
Estimated_Total_Intracranial_Volume  4.947e-04  3.776e-03   0.131  0.89577
Atlas_Scaling_Factor                -2.866e+00  4.831e+00  -0.593  0.55296
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1280.55  on 950  degrees of freedom
Residual deviance:  597.45  on 942  degrees of freedom
AIC: 615.45

Number of Fisher Scoring iterations: 7
```

```
> library(broom)
>
> tidy(log_reg)
# A tibble: 9 × 5
  term                               estimate std.error statistic  p.value
  <chr>                                 <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)                        -31.7       11.5       -2.77  5.66e- 3
2 Gender_CategoryMale                 -1.89       0.281     -6.72  1.84e-11
3 Age_GroupAge Between 73 and 85       0.373      0.237      1.57  1.16e- 1
4 Age_GroupAge Between 86 and 98       1.94       0.429      4.52  6.07e- 6
5 Years_of_Education                   0.177      0.0614     2.88  3.98e- 3
6 Socioeconomic_Status                 0.344      0.146      2.36  1.84e- 2
7 Mini_Mental_State_Examination_Score  1.13       0.0894    12.7   7.15e-37
8 Estimated_Total_Intracranial_Volume  0.000495   0.00378    0.131 8.96e- 1
9 Atlas_Scaling_Factor                -2.87       4.83      -0.593 5.53e- 1
>
```

Characteristic	**OR**[1]	**95% CI**[1]	**p-value**
Gender_Category			
Female	—	—	
Male	0.15	0.09, 0.26	<0.001
Age_Group			
Age Between 60 and 72	—	—	
Age Between 73 and 85	1.45	0.91, 2.31	0.12
Age Between 86 and 98	6.98	3.09, 16.7	<0.001
Years_of_Education	1.19	1.06, 1.35	0.004
Socioeconomic_Status	1.41	1.06, 1.89	0.018
Mini_Mental_State_Examination_Score	3.11	2.63, 3.74	<0.001
Estimated_Total_Intracranial_Volume	1.00	0.99, 1.01	0.9
Atlas_Scaling_Factor	0.06	0.00, 844	0.6

[1] OR = Odds Ratio, CI = Confidence Interval

b) Create a confusion matrix and explain how well the model is classifying who has Dementia.

Solution:

After creating the Logistic Regression model using the train dataset, we use that model to run on the test dataset. We have used a 80-20 ratio for train and test dataset. We have 191 rows in the test dataset.

Below is the Confusion Matrix for the testing dataset:

- The accuracy of the model is 87.96%. It means that the model is able to accurately classify patients who actually have Alzheimer and who do not have Alzheimer.
- The balanced accuracy of the model is 87.17% which means that of the data is imbalanced when the actual target group having Alzheimer is in less proportion.
- The sensitivity of the model is 83.12% which means that the model is able to classify the patient having Alzheimer when in reality the patient has Alzheimer and it is accurate close to 83% of the times.
- The miss rate of the model is 1 - sensitivity = 100 - 83.12 = 16.88%, which means that the model goes wrong close to 17% of the times and classifies a patient having No Alzheimer when in reality the patient has Alzheimer.
- The specificity of the model is 91.23% which means that the model is able to classify the patient having No Alzheimer when in reality the patient has No Alzheimer and it is accurate close to 91% of the times.

```
> # Creating a model on training dataset
> train$Dementia<- as.factor(train$Dementia)
>
> log_reg_train = train(
+   form = Dementia ~ Gender_Category + Age_Group + Years_of_Education +
+                   Socioeconomic_Status + Mini_Mental_State_Examination_Score +
+                   Estimated_Total_Intracranial_Volume + Atlas_Scaling_Factor,
+   data = train,
+   method = "glm",
+   family = "binomial"
+ )
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
>
> # Running the model on the test dataset and getting the Confusion Matrix
> confusionMatrix(predict(log_reg_train, test), as.factor(test$Dementia))
Confusion Matrix and Statistics

              Reference
Prediction     Alzheimer No Alzheimer
  Alzheimer           64           10
  No Alzheimer        13          104

               Accuracy : 0.8796
                 95% CI : (0.8248, 0.9221)
    No Information Rate : 0.5969
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7482

 Mcnemar's Test P-Value : 0.6767

            Sensitivity : 0.8312
            Specificity : 0.9123
         Pos Pred Value : 0.8649
         Neg Pred Value : 0.8889
             Prevalence : 0.4031
         Detection Rate : 0.3351
   Detection Prevalence : 0.3874
      Balanced Accuracy : 0.8717

       'Positive' Class : Alzheimer
```

c) Create an ROC curve and calculate the c-statistic (auc). What does this mean about the model?

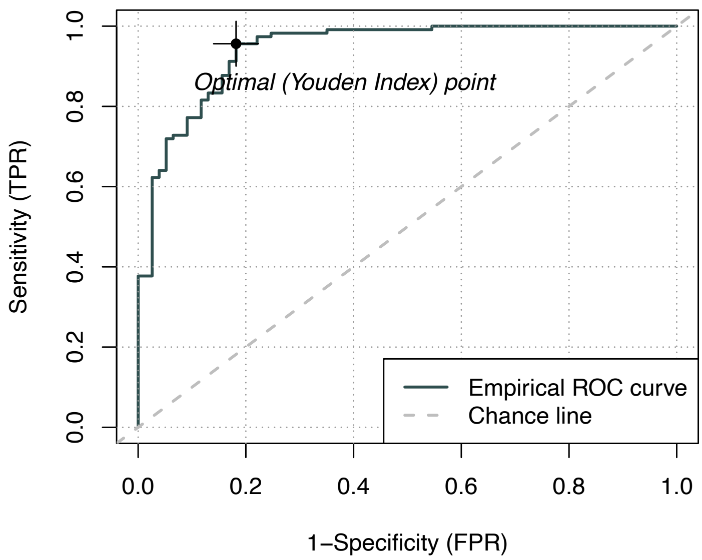
<u>Solution:</u>

The ROC curve is the curve of the True Positive Rate (TPR) to the False Positive Rate (FPR) for the model and it tells us how well the model is suitable for classifying.

The ROC curve for the Logistic Regression model is as follows:

The model can be used to predict the probability of a point belonging to a particular class based on the TPR and the FPR values. In binary classification, there may be cases when the threshold for deciding whether a coin belongs to a specific class may not always be 0.5. The threshold values may change. The ROC curve helps us to estimate the best threshold and metric for the model to provide the best classification.
C-statistic is the Area under the ROC Curve and it tells us how well the model is able to classify whether the data point belongs to positive or negative class. Higher the AUC, better is the ability of the model to distinguish between the positive and negative classes.

In our case, the c-statistic value is 0.942. This means that the model is great at distinguishing the patients as having Alzheimer and No Alzheimer. This is because the model is able to detect more true positives and true negatives than false positives and false negatives.




```
> summary(ROCit_obj)

 Method used: empirical
 Number of positive(s): 114
 Number of negative(s): 77
 Area under curve: 0.9422
```

d) What are the differences between the information in part a and part b?

Solution:

In part a of the question, we have run a Logistic Regression model and here we are trying to understand from the healthcare point of view as to which of the input variables to the model have a significant contribution in determining of a person can have Alzheimer or not. In this case, we use the Odds Ratio to see which are important. Out of the variables which are significant, the ones having Odds Ratio > 1 are risk factors and those with Odds Ratio < 1 are preventative factors. We understand that Males are at a higher risk of suffering from Alzheimer as compared to females. This gives us an overall idea as to which factors can be detrimental to the patients condition and which can lower the chances of a person having Alzheimer.

In part b of the question, we have run a Logistic Regression model and here we are trying to understand from the statistics or machine learning point of view as to how good the generated model is at identifying whether a patient has Alzheimer or not. We understand the accuracy and performance of the model in terms of correctly classifying the patients' condition, the miss rate when it incorrectly classifies a patient being fine when in reality the patient has Alzheimer and the cases when it accurately identifies a patient as Alzheimer when in reality the patient has Alzheimer.

e) How does this model differ from the linear discriminate analysis you ran in Assignment 7?

Solution:

When we compare the Linear Discriminant Analysis technique and Logistic Regression techniques, both are used to classify the output dependent variables but there are a few differences.

Linear Discriminant Analysis can only work when all the input independent variables are numeric in nature whereas the Logistic Regression can work even when the input independent variables are numeric, categorical or binary in nature.
In our case when we used Linear Discriminant Analysis, we had all the variables as numeric but in Logistic Regression, we converted gender and age to categories and it worked.

Linear Discriminant Analysis has its roots in linear regression and assumes multivariate normality, linearity, multicollinearity and homoscedasticity. But such conditions are not required for Logistic Regression.

Now, let us compare the Linear Discriminant Analysis and Logistic Regression models.

```
>
> # Running the classifier on test data
> pred_TT_test = predict(alzheimerLDA_TT, newdata=test[,c(1:10)])$class
> table_TT_test<-table(pred_TT_test, test$Dementia)
> table_TT_test

pred_TT_test   Alzheimer No Alzheimer
  Alzheimer           78            1
  No Alzheimer         1          110
> confusionMatrix(table_TT_test)
Confusion Matrix and Statistics


pred_TT_test   Alzheimer No Alzheimer
  Alzheimer           78            1
  No Alzheimer         1          110

               Accuracy : 0.9895
                 95% CI : (0.9625, 0.9987)
    No Information Rate : 0.5842
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9783

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9873
            Specificity : 0.9910
         Pos Pred Value : 0.9873
         Neg Pred Value : 0.9910
             Prevalence : 0.4158
         Detection Rate : 0.4105
   Detection Prevalence : 0.4158
      Balanced Accuracy : 0.9892

       'Positive' Class : Alzheimer
```

In the Linear Discriminant Analysis technique, we get an overall accuracy of 98.95%. It means the model is able to classify the patient having Alzheimer or not to a great extent correctly. The sensitivity of 98.73% tells us correctly that a patient has Alzheimer when in reality they have Alzheimer. The miss rate is 100 - 98.73 = 1.57% which means that the model is good and classifies a Alzheimer patient as No Alzheimer is approximately 2% of the times. The specificity is also 99.1% which indicates that it correctly labels a patient having No Alzheimer when they actually are fine.

```
Confusion Matrix and Statistics

              Reference
Prediction     Alzheimer No Alzheimer
  Alzheimer           64           10
  No Alzheimer        13          104

               Accuracy : 0.8796
                 95% CI : (0.8248, 0.9221)
    No Information Rate : 0.5969
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7482

 Mcnemar's Test P-Value : 0.6767

            Sensitivity : 0.8312
            Specificity : 0.9123
         Pos Pred Value : 0.8649
         Neg Pred Value : 0.8889
             Prevalence : 0.4031
         Detection Rate : 0.3351
   Detection Prevalence : 0.3874
      Balanced Accuracy : 0.8717

       'Positive' Class : Alzheimer
```

In the Logistic Regression technique, we get an overall accuracy of 87.96%. It means the model is able to classify the patient having Alzheimer or not to a good extent correctly but not as efficiently as the LDA. The sensitivity of 83.12% tells us correctly that a patient has Alzheimer when in reality they have Alzheimer. The miss rate is 100 - 83.12 = 16.88% which means that the model is not very good and classifies a Alzheimer patient as No Alzheimer is approximately 17% of the times. The specificity is also 91.23% which indicates that it correctly labels a patient having No Alzheimer when they actually are fine.

In conclusion, the Linear Discriminant Analysis model is better at classifying the patient as having Alzheimer or not as compared to the Linear Regression model.