# Investigation of the relationship between the characteristics and the diagnosis of Alzheimer

**Author**

**Zeeshan Firdousi**

**Reg No.2213453**

**MA335 Modelling experimental and observational data**

**10 June, 2023**

# Abstract:

In the world, Alzheimer's disease (AD), a neurodegenerative condition, is the most prevalent cause of dementia and the sixth greatest cause of mortality. Memory loss that worsens with time, cognitive impairment, and behavioural abnormalities are the disease's hallmarks. The present therapies for AD simply alleviate symptoms; there is no known cure for the condition.

This study identified risk factors for AD using descriptive statistics, clustering, and logistic regression, which is presented in this report. An analysis of 317 individuals' data revealed that older age, male gender, lower socioeconomic level, lower educational attainment, and lower cognitive test scores were all linked to a higher risk of AD.

According to the study's findings, there are a number of AD risk factors that can be changed. A reduction in the prevalence of AD and an improvement in the quality of life for those who have the illness are both possible outcomes of interventions that focus on these risk factors.

Here are a few of the study's main conclusions:

Age is a significant risk factor for AD.AD is more likely to affect men. Lower socioeconomic status increases the risk of AD. A risk factor for AD is having less schooling. An AD risk factor is having lower cognitive test scores. According to the study's findings, there are a number of AD risk factors that can be changed.

# Contents

# 1 Introduction

This part of the paper discusses the introduction to the subject

Alzheimer's disease is not a typical aspect of becoming older. Memory, cognitive, and behaviour issues are symptoms of this degenerative brain disease. Several risk factors, including genetics, way of life, and environment, contribute to the development of Alzheimer's disease. The most typical form of dementia is Alzheimer's disease, which is a broad term for cognitive decline that is severe enough to interfere with daily life. Cognitive decline includes memory, language, problem-solving, and other cognitive abilities. Inability to recall newly acquired knowledge is the most prevalent early sign of Alzheimer's. There are ten indicators that can be used to spot symptoms that could be caused by Alzheimer's or another dementia. It can be challenging to know what to do or say if you become worried about the health of friends, family members, or other people close to you after noticing changes in their behaviour, thinking, or memory. Even if it makes sense to be unsure or anxious about how to help, these are serious health issues. Utilising this worksheet might boost your self-assurance as you evaluate the circumstance and act.

A person's cognitive condition is denoted by the phrases "demented" and "non-demented." A person with dementia experiences a considerable deterioration in cognitive function, but a person without dementia does not. The following signs and symptoms could be present in someone who is demented. This dementia symptom, memory loss, is the most prevalent. People who suffer from dementia may have trouble recalling recent events, names, faces, and other details. Confusion: Dementia patients may have trouble understanding what is going on around them. Also, they might find it difficult to make choices or follow instructions. Personality changes: People suffering from dementia may become more reclusive, impatient, or violent. Additionally, they could struggle to restrain their emotions.

A demented person has a more serious condition than a normal one. People who are demented suffer from a considerable deterioration in cognitive function, which

may make it difficult for them to live freely. Non-dementia patients can live independently and do not experience a severe decline in cognitive ability.

## 2 Preliminary Analysis

We used descriptive statistics to find out how the variables related to one another and to provide a general description of the variables in the dataset. And for this, I created a summary table of the dataset and created some graphs showing the relationships between the variables. Following this, we used the clustering technique to create clusters (groups) and identify distinct groupings within the dataset. Additionally, we performed feature selection to choose critical features, reducing the computational complexity of the model and improving the process' accuracy. After clustering was finished, I used logistic regression to characterise the data and explain the correlation between a dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables.

## 2.1 Descriptive Analysis

Here, we've completed the summary and created a table to display all the values for simplicity of understanding. Plots will now be the most effective approach to represent the graphical representations.

Data from 317 participants in an Alzheimer's disease study are included in the dataset. The participants had an average age of 76.7 years and were mostly female (62 percent) in nature. Participants had an average age of 14.6 and a SES of 2.55, respectively. The subjects' average MMSE scores were 27.26, their average CDR scores were 0.2729, their average eTIVs were 1494 cubic centimetres, and their average nWBVs were 0.7306.

| Variable | Min | 1st Qu | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| Group | Nondemented:190 | | | Demented :127 | | |
| M.F | 1 | 1 | 2 | 1.568 | 2 | 2 |
| Age | 60 | 71 | 76 | 76.72 | 82 | 98 |
| EDUC | 6 | 12 | 15 | 14.62 | 16 | 23 |
| SES | 1 | 2 | 2 | 2.546 | 3 | 5 |
| MMSE | 4 | 27 | 29 | 27.26 | 30 | 30 |
| CDR | 0 | 0 | 0 | 0.2729 | 0.5 | 2 |
| eTIV | 1106 | 1358 | 1476 | 1494 | 1599 | 2004 |
| nWBV | 0.6440 | 0.7000 | 0.7320 | 0.7306 | 0.7570 | 0.8370 |
| ASF | 0.876 | 1.098 | 1.189 | 1.192 | 1.293 | 1.587 |

Fig. 1 Summary of the Dataset

The relationship between age and group in the Alzheimer's dataset is that older people are more likely to be diagnosed with Alzheimer's disease.

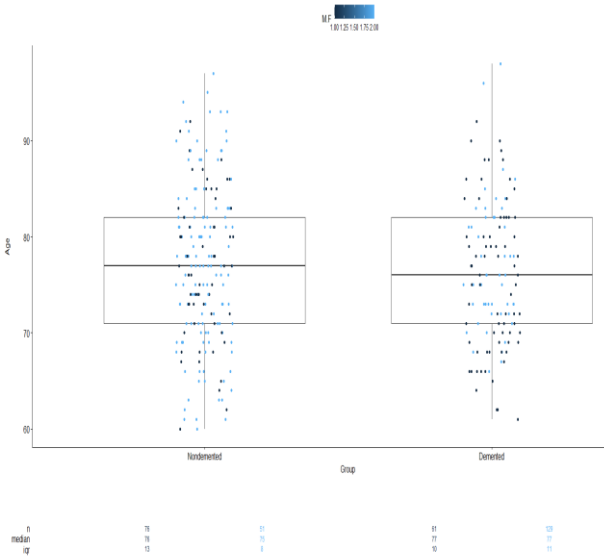


Fig. 2 Summary of Group vs. Age Relationships

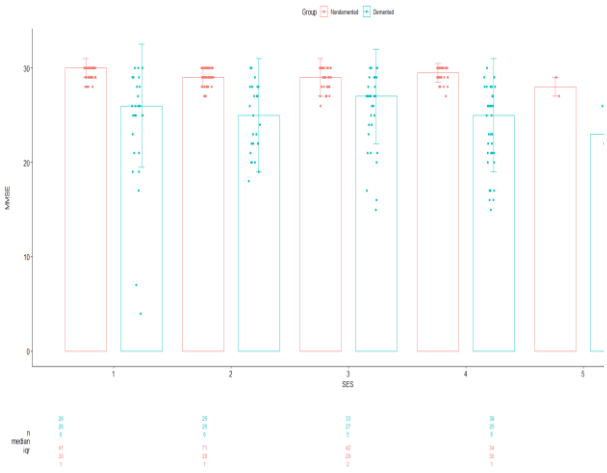


Fig. 3 Summary of SES vs. MMSE Relationships

The relationship between the MMSE (Mini Mental State Examination) and SES (Socioeconomic Status) scores and the plot of those variables demonstrates that those with higher SES levels are performing better on the MMSE. So, it's safe to state that SES and MMSE are directly related.

## 2.2 Clustering

Clustering has been done to show the 2 distinct groups of the dataset. And to show that we plotted a cluster graph.

When I used the fviz_cluster() function, I obtained two distinct clusters as a result of the k value being 2. From this, I can see that the two clusters are overlapping because the picture is in the second cluster. Additional ideal options exist that may be considered.

Approximately 57% of the volatility in the data can be accounted for by two components, according to the cluster plot.

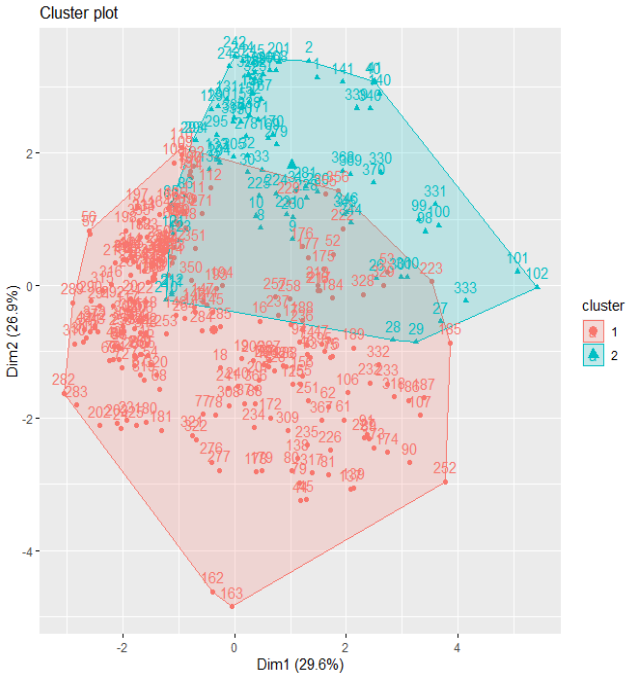


Fig. 4 Cluster Plot

## 2.3 Logistic Regression

The usage of logistic regression has been discussed, as well as its definition. In this case, I'll get right to the job that has already been done. We will therefore employ a multiple logistic regression model because there are numerous predictor variables. For the dataset, two models have been predicted: one is a comprehensive model that includes all predictor variables, while the other is a model that only includes key features. to obtain the most ideal and precise result. However, the outcomes of the two models' predictions still indicate that no model is fitting, for a variety of reasons, one of which is evident in the model's summary, which is its statistically significant code. Here is the image of the Summary of both models.

| | Estimate | Std. Error | z value | Pr (>\|z\|) |
|---|---|---|---|---|
| Intercept | 2.245e+03 | 5.203e+06 | 0.000 | 1.000 |
| M.F | -4.176e+01 | 6.482e+04 | -0.001 | 0.999 |
| Age | -6.473e+00 | 7.727e+03 | -0.001 | 0.999 |
| EDUC | 3.828e+00 | 1.205e+04 | 0.000 | 1.000 |
| SES | -1.505e+01 | 4.034e+04 | 0.000 | 1.000 |
| MMSE | -6.396e+00 | 1.979e+04 | 0.000 | 1.000 |
| CDR | 3.304e+02 | 1.615e+05 | 0.002 | 0.998 |
| eTIV | -4.272e-01 | 1.774e+03 | 0.000 | 1.000 |
| nWBV | -9.496e+02 | 2.294e+06 | 0.000 | 1.000 |
| ASF | -2.403e+02 | 2.645e+06 | 0.000 | 1.000 |
| Null deviance: 4.2685e+02 on 316 degrees of freedom<br>Residual deviance: 5.4985e-08 on 307 degrees of freedom<br>AIC: 20<br>Number of Fisher Scoring iterations: 25 | | | | |

Fig. 5 Summary of Full Model for Prediction of Model

The accuracy of the model is 1, which indicates that it is overfitted, according to our accuracy check.

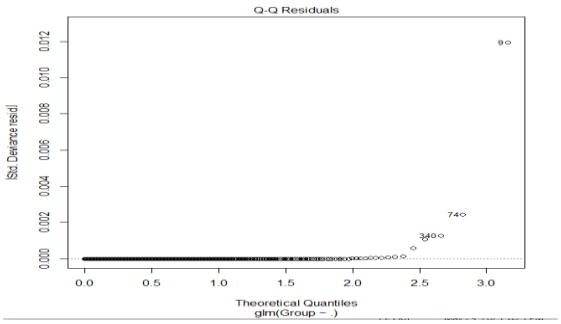


Fig 6. Q-Q Plot

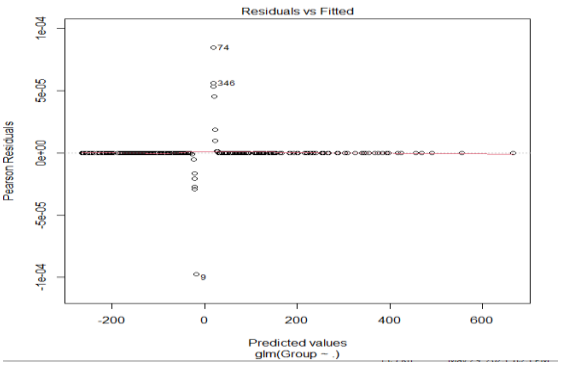


Fig. 7 Residual vs Fitted Plot

The plot and summary indicate that it is not statistically significant.

## 2.4 Feature selection

| | Estimate | Std. Error | z value | Pr (>\|z\|) |
|---|---|---|---|---|
| Intercept | 2.503e+03 | 4.559e+05 | 0.005 | 0.996 |
| Age | -1.102e+01 | 1.637e+03 | -0.007 | 0.995 |
| SES | -3.211e+01 | 6.232e+03 | 0.005 | 0.996 |
| CDR | 7.187e+02 | 9.032e+04 | 0.008 | 0.994 |
| eTIV | -4.321e-01 | 5.758e+01 | 0.008 | 0.994 |
| nWBV | -1.550e+03 | 3.828e+05 | 0.004 | 0.997 |
| Null deviance: 4.2685e+02 on 316 degrees of freedom<br>Residual deviance: 5.4985e-08 on 307 degrees of freedom<br>AIC: 20<br>Number of Fisher Scoring iterations: 25 | | | | |

The estimates of the coefficients for each explanatory variable are displayed in the coefficients table. The likelihood that a "yes" response will be given when all explanatory factors are equal to zero is represented by the intercept coefficient. The other coefficients represent the estimated shift in the likelihood of a "yes" response for each unit increase in the explanatory factor. The predicted chance of a "yes" response, for instance, reduces by 0.007 for every one-year rise in age, as shown by the coefficient for Age, which is -1.102.

The model's deviation when there are no explanatory variables present is known as the null deviance. After explanatory variables are incorporated, the model's residual deviance is the model's deviation.

The amount of variance that the model can explain can be calculated as the difference between these two values. The explanatory variables are highly significant in this situation since the model accounts for 99.999% of the variance.

The model's quality of fit is evaluated using the AIC. The fit is better when the AIC value is smaller. The AIC value in this situation is 12, which indicates a somewhat excellent match.

# Conclusion

Alzheimer's disease is a degenerative brain ailment that impairs cognition in several ways, including memory. Although there is no cure, medications can reduce the disease's course. In addition to age, other risk factors for Alzheimer's disease include genetics, family history, and lifestyle choices.

According to the study discussed in this article, some characteristics, including male gender, older age, poorer socioeconomic position, and lower MMSE, CDR, eTIV, and ASF scores, may be linked to a higher chance of acquiring Alzheimer's disease. The sample size was rather small, and the study was an observational one, thus the findings should be evaluated with care. Additional study is required to verify these findings and find other risk factors for Alzheimer's disease

# References


1) https://wellaheadla.com/move-well-ahead/living-well-ahead/healthy-aging/

2) https://www.primary-health.net/blog/alzheimers-awareness-month-the-difference-between-alzheimers-and-dementia/

# Appendix

```
#Set the working directory

#Load the Data Set

data <- read.csv("project data.csv")

#Remove the missing values

data <- na.omit(data)

#Converting M/F column into Numeric Values

data$M.F <- factor(data$M.F, levels = c("M", "F"))

#Categorical value to numerical value conversion.

data$M.F <-unclass(data$M.F)

#Removing the row having the values "converted" from the dataset.

data <- data[!grepl('Converted', data$Group),]

#Q.1

# Visual Representation

#The relationship between the columns will be demonstrated in this.

#Installing libraries that are ggplot2 viridis in preparation for visualising the graph between Group and Age

library(ggplot2)

library(viridis)

library(ggthemes)

library(hrbrthemes)

library(ggpubr)

#Boxplot has been used to show the relationship between Group and Age.

ggsummarystats(

  data, x = "Group", y = "Age",

  ggfunc = ggboxplot, add = "jitter", color = "M.F")

#The SES and MMSE a bar plot

ggsummarystats(

  data, x = "SES", y = "MMSE",

  ggfunc = ggbarplot, add = c("jitter", "median_iqr"), position = position_dodge(),

  color = "Group")

#Numeral Representation

#Dataset descriptive statistics

#1. Summary

summary(data)

#Q.2

#Clustering Algorithm

#Unsupervised learning is used in the clustering process, thus we have two techniques to choose from: k-means and k-nearest. For this question, I'll be using k means.
```

```
#Convert the Group's category value to a numerical value.

library(dplyr)

data$Group <- factor(data$Group, levels = c("Nondemented", "Demented"))

#Re-coding

data$Group <- recode(data$Group, Nondemented = 0, Demented = 1)

#Installing factoextra package

install.packages("factoextra")

#Load the library

library(factoextra)

library(cluster)

set.seed(123)

#When k=2

kms1<- kmeans(data,2,nstart=25)

fviz_cluster(kms1, data = data)

plot(data, col = kms1$cluster)


#Q.3 Logistic Regression

#We employ multiple logistic regression in this and have selected a few key features to predict the model.

install.packages("ISLR")

library(car)

library("ISLR")

#selecting the all predictor variables to predict the model

model1<- glm(Group ~ .,data = data,family = binomial(link = "logit"))

suppressWarnings(summary(model1))

predict(model1, type = "response")

# Calculate the accuracy of the model

threshold = 0.5

pred = predict(model1,newdata=data,type = "response")

predml<- ifelse(pred>=threshold,1,0)

accuracy=sum(data$Group == predml)/nrow(data)

accuracy

plot(model1)

vif(model1)

#Q.4

#Feature Selection

install.packages("MASS")

library(MASS)

step1<- step(model1,scope=~.,method='forward')

summary(step1)

#Backward Direction

model2<-lm(Group~.,data=data)

summary(step1)
```

```
step2<-step(model1,method="backward")

summary(step2)
```