Research Project
Relationship Between Education and Alzheimer's Severity

Cesare Bergossi, Giulia Pezzani

January 2023

1 Introduction

Alzheimer's disease is a progressive neurological disorder that affects memory, thinking, and behavior. It is the most common cause of dementia in older adults, and the number of people affected by Alzheimer's is expected to increase as the population ages.

One potential risk factor for Alzheimer's is education level. Some studies have suggested that individuals with higher levels of education may have a lower risk of developing Alzheimer's, while others have found no relationship or even an increased risk for those with higher education. The relationship between education and Alzheimer's risk is not well understood, and more research is needed to clarify this relationship.

In this study, we aim to examine the relationship between education and the risk of developing Alzheimer's disease in older adults, measured by the CDR (Clinical Dementia Rating). We will use a large, nationally representative dataset to analyze the odds of having Alzheimer's disease for individuals with different levels of education, controlling for other potential confounding factors such as age, sex, and overall health status. Our findings will contribute to the understanding of the role of education in the development of Alzheimer's disease and may inform the development of interventions to prevent or delay the onset of the disease.

2 Dataset Exploration

2.1 Description

The dataset we have chosen for this project is part of the datasets collection of the website Kaggle. It includes clinical and physical data on patients with Alzheimer's disease, such as age, gender, education level and cognitive test scores. Specifically, they are labelled as:

- **M.F** - Patient's gender;
- **Age** - Patient's age;
- **EDUC** - Years of education;
- **SES** - Socioeconomic status (1-5);
- **MMSE** - Mini Mental State Examination (0-30), a tool used to assess cognitive function in patients with dementia and other neurological conditions;

- **CDR** - Clinical Dementia Rating (0-3), used to assess the severity of dementia in a patient;
- **eTIV** - Estimated Total Intracranial Volume;
- **nWBV** - Normalized Whole Brain Volume;
- **ASF** - Atlas Scaling Factor, the volume-scaling factor required to match each individual to the atlas target.

2.2 Data Cleaning

Before analyzing the data, we will perform some data cleaning steps to ensure its quality. First, we check for missing values and substitute them with the average. Then, we model dummy variables for the "M.F" column. Lastly, we will check for outliers after representing our data graphically.

2.3 Visualization

Next, we visually explore our dataset using different plots; we plot different variables (age, years of education, socioeconomic status, etc.) against Clinical Dementia Rating. Doing this, we are able to identify potential outliers, data points that are significantly different from the rest of the data. Therefore, we exclude them from our analysis, since they could influence the results and conduct to misleading conclusions.
From these scatter plots, we begin to have an idea of possible connections between singular variables and CDR: for instance, MMSE and physical conditions (brain and intracranial volume) show some sort of interdependence, while other factors like age and years of education did not exhibit particular patterns.
Additionally, we depict in two boxplots the results in the MMSE (Mini Mental State Examination) of both men and women, to identify potential differences: we noticed slightly lower scores on the men side, although we have not yet showed whether there is a correlation between this examination and dementia.
We chose not to plot different genders against CDR, since the variety of the values taken by the latter is too little to note a significant pattern between male and female. Further research could analyze eventual visual correlations through a 3D plot which also takes into account the frequency of each CDR result with respect to gender.

3 Results

3.1 Linear Regression

After this preliminary data exploration, we proceed by modelling a multivariate linear regression. The main scope of this research project is to find out if there is some sort of correlation between education level and Alzheimer's severity; however, in our analysis we also want to take into account other predictors for CDR, which might influence the outcome.

Analyzing the outcome of this regression, we can immediately notice that some of the predictors (eTIV, ASF) are not significant since their p-values are larger than 0.05. Further observation tells us, moreover, that the adjusted R-squared value is not sufficiently high, so that the contribution of each predictor to the overall model fit is not strong enough. This index is, indeed, a measure of the goodness of fit of the model that adjusts for the number of predictors.

For these reasons, after performing this linear regression, we will select a specific model, first using a test method (Step-Down) and then using a penalty method (LASSO). Model selection is particularly useful because it allows us to identify the most appropriate model for our data and can improve the accuracy and interpretability of the results, other than possibly avoiding the risk of overfitting. Finally, after comparing these two methods, we will check if the assumptions of normality and homoscedasticity for the residuals are satisfied.

3.2 Step-Down

Through Step-Down method we select a subset of variables for our multiple linear regression model. It works by starting with all the variables in the model and then iteratively removing the least significant variable until all the remaining ones are significant.
By this, we mean that the iteration will stop when all $H_0 : \beta_i = 0$ are rejected (separately) for every parameter β_i through a t-test at level $\alpha = 0.05$. If that is not the case, we remove the predictor with highest p-value and repeat the testing.
Luckily, we already have the p-values from the linear regression, so the iteration is much less tedious.

3.3 LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) is a regularization method for linear regression that uses an L1 penalty to encourage sparse solutions, i.e., solutions with many coefficients equal to zero. This can be useful for model selection, as it can automatically identify a subset of significant predictors.
We first select an appropriate value for lambda through a 10-fold cross validation (an arbitrary value of lambda would lead to potential errors, such as overfitting or exclusion of optimal models). After finding lambda, we convert the data to a matrix and fit the LASSO model using the glmnet function, setting the lambda parameter equal to the one we just found. We then extract the coefficients from the model and select only the non-zero ones, which represent the significant covariates.

3.4 Comparison of Model Selection Methods

The iterative t-test for the step-down method ran once, eliminating the predictor eTIV, leaving us with 7 covariates. The multiple R-squared and the adjusted R-squared indices (respectively, 0.5315 and 0.5223) tell us, however, that there is not such an evident correlation between the covariates and the response in this linear model, as they are almost equal to the ones obtained through the multiple linear regression.
We now perform an F-test to compare the step-down model with just the intercept, to check whether this model is better than none; we use the anova() function in R, which performs an F-test to compare the fits of these two models. This test returns a p-value $< 2.2e^{-16}$ (indeed much smaller than 0.05), so we can conclude that the step-down model is at least better than no model.

The LASSO method, instead, keeps all the predictors (as all of them have non-zero coefficient in the LASSO model), so we keep our original multivariate linear regression with 8 covariates (therefore, also the original R-squared values 0.5315 and 0.521, which showed a weak correlation).

Another observation is that both methods successfully avoided overfitting, as the difference between the multiple R squared and the adjusted R squared is minimal.

3.5 Normality and Homoscedasticity of Residuals

One important step in this part of the project is to examine the residuals in our multiple linear regression to assess the model fit and to check for assumptions that may have been violated. In order to do so, we perform two kinds of inspections:

- Graphical methods - We plot the values of the residuals we found through the Step Down method and the LASSO.
- Normality tests - We perform a Kolmogorov-Smirnov and a Shapiro-Wilk test, which check for normality.

Let us consider the Model Selection Methods that we just used:

1. Step-Down - By plotting the residuals from the step-down method against residuals from a Gaussian distribution in a Q-Q Plot, it can be noticed that there is a significant deviation from normality. Moreover, we performed two normality tests (Kolmogorov-Smirnov and Shapiro-Wilk) and both of them resulted into a divergence from a normal distribution, with very low p-values.
   However, regarding homoscedasticity, a scatter plot of the residuals shows that the distribution of variance remains pretty much the same across the fitted values, so that we can retain this assumption.

2. LASSO - The results for normality in the LASSO method are very similar to the step-down, except for the fact that the Q-Q Plot shows an even stronger deviation from normality, and the p-value obtained from the Shapiro-Wilk test is much smaller. Note that we did not run a Kolmogorov-Smirnov test, as it requires the data not to contain ties, which occurs when multiple observations have the same value.
   Even for the LASSO method, we can assert that the homoscedasticity assumption was attained.

4 Hypothesis testing

Back to the focus of our research project, through our linear regression and consequent model selection, we observed that the variable "Years of Education" was not discarded; at the same time, however, the R-squared values were too low to state that there is a certain correlation between Education and Alzheimer's severity.
This is why we decided to perform a hypothesis test to further investigate if education represents a meaningful factor when analyzing Clinical Dementia Rating.

We will perform a t-test to compare the means of CDR scores for two groups: those with fewer years of education (under 15) and those with more years of education (over 15). Analyzing the normality of the CDR in the two samples with a Shapiro-Wilk test, we can notice that the two groups are not normally distributed (due to very low p-values). Even if their variances are very close one to the other (0.1225029 and 0.1041201), the fact that normality is not attained leads us to opt for an asymptotic two sample t-test. Even the number of observations in both groups, respectively 185 and 180 (quite large), justifies our choice.
We will formulate our hypothesis testing as follows

$$H_0 : \mu_1 < \mu_2 \text{ vs. } H_1 : \mu_1 > \mu_2$$

where μ_1 is the mean of the group with less year of education and μ_2 is the mean of the other group. By running the t-test, we find a T statistic of 2.8199, which leads us to a p-value = 0.002534; as the latter is lower than the significance level 0.05, we reject the null hypothesis, meaning that μ_1 is indeed significantly greater than μ_2.
We can then conclude that there is firm evidence that the level of dementia in patients with less years of education is higher than the one in patients with more years of education.

5 Conclusion and limitations

Throughout this study, we modeled our dataset and looked for patterns, focusing on the response variable CDR (Clinical Dementia Rating), which efficiently describes Alzheimer's severity in patients.
We used a series of statistical tools, first starting with visual inspections through several plots, from which we got an idea of potential patterns in the relation between variables and CDR, even if our target predictor (EDUC) did not show specific signs of connection at first sight.
We then proceeded with computationally analyzing the correlation between CDR and all of the other factors through a multivariate linear regression, which already showed us non-significance of some predictors and in general a poor correlation.
This was later confirmed through some model selection techniques: the step-down method discarded the covariate that we were expecting and it showed even weaker correlation between the predictors and the response variables, while the LASSO chose the full model as optimal.
It is important to note that in both cases, the education level was kept, however this analysis was not enough to state whether this factor is impactful in the prediction of Alzheimer's intensity.
This is what led us to additional research through an asymptotic T-test; this eventually revealed that the severity of the disease in patients with a higher level of education was lower.

Even if this result is fairly satisfying for the purpose of this research, there still are some constraints. One potential limitation of this study is that the sample may not be representative of the entire adult population. For instance, this dataset oversamples individuals who are already retired or nearing retirement, so the results may not be generalizable to younger adults or those who are still working.
Another obstacle is the fact that we did not check if predictor variables are related one to the other (just think about the difference in MMSE results based on gender, which we showed in a boxplot): more research should be done in order to assert this.

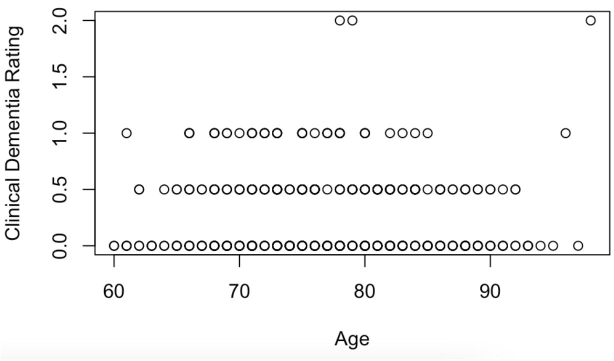
This research project showed important consequences based on the Education Level, however this medical disease still has a lot of unknowns; for example, it should be asked whether CDR score can be predicted using a combination of demographic, genetic and imaging measures, to which our dataset was completely blind.

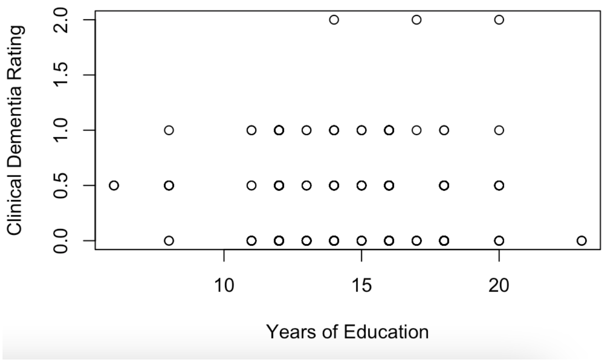
References

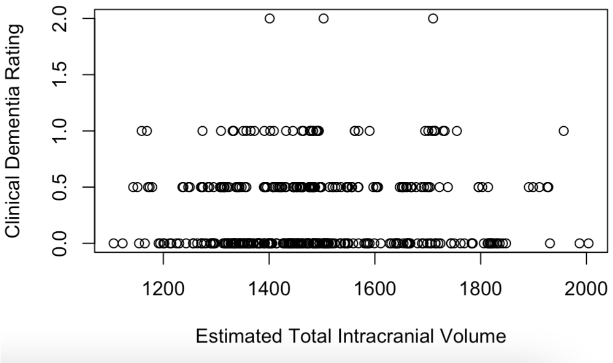
- ”An Introduction to Mathematical Statistics” (Fetsje Bijma, Marianne Jonker etc.)
- Link to the dataset: *https://www.kaggle.com/datasets/brsdincer/alzheimer-features*
- ”Geeks for Geeks” website *https://www.geeksforgeeks.org*

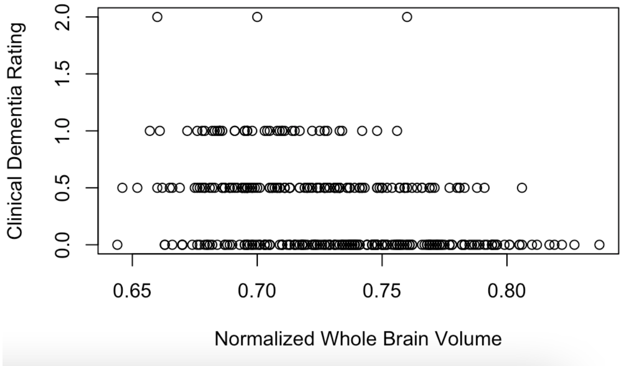
Appendix

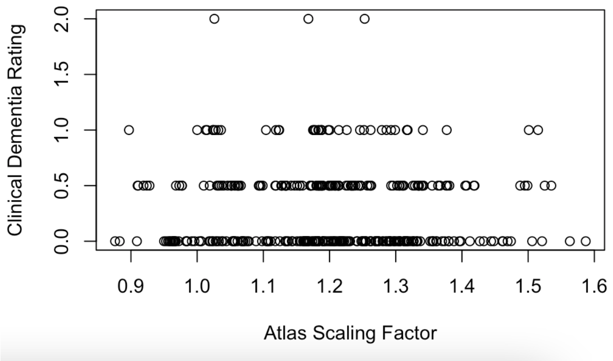
5.1 Visual Exploration

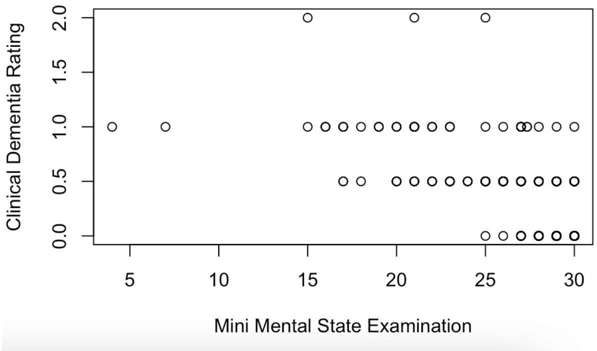


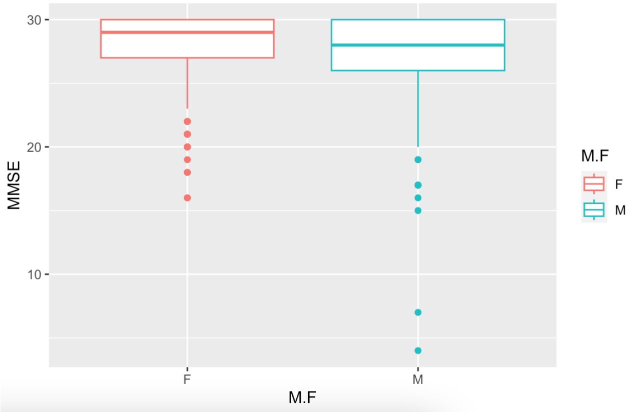




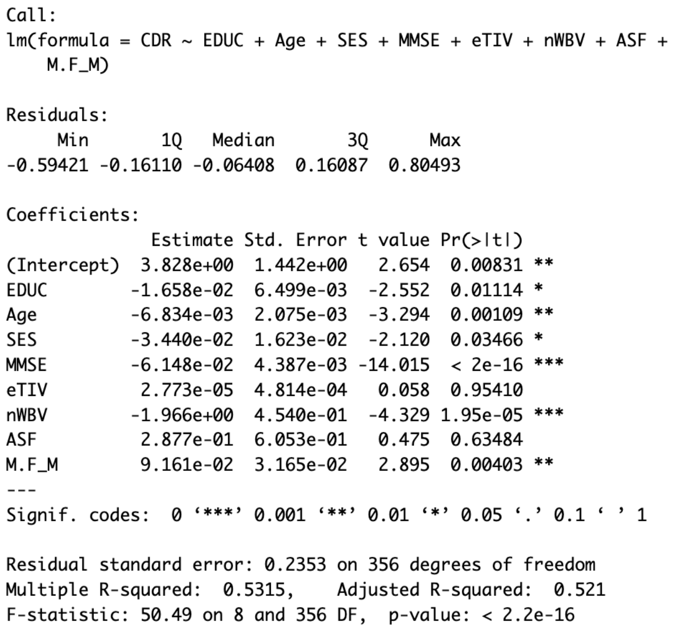






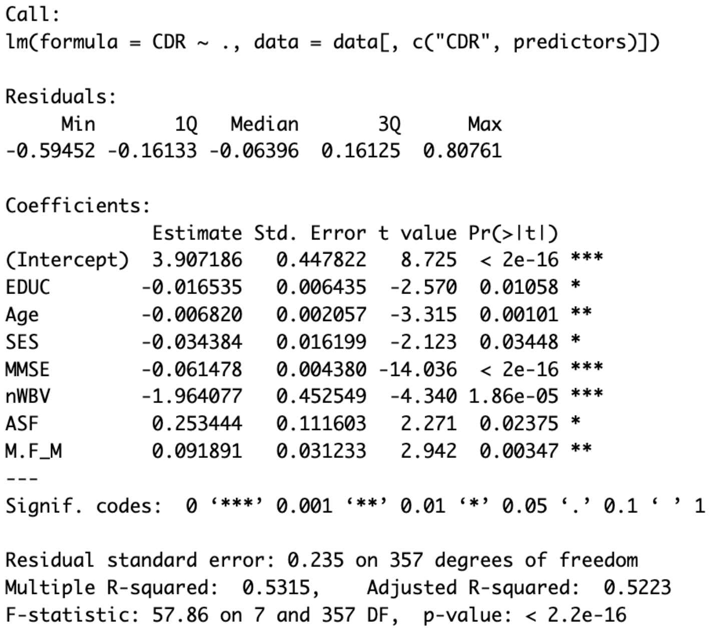


5.2 R Output - Linear Regression



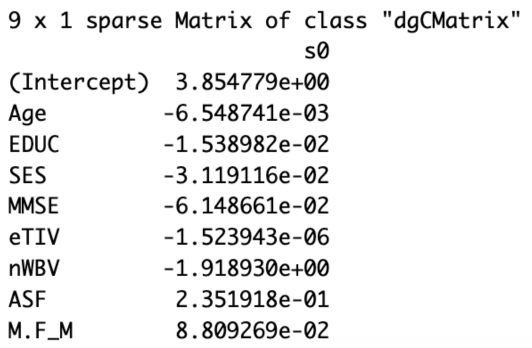
```
Call:
lm(formula = CDR ~ EDUC + Age + SES + MMSE + eTIV + nWBV + ASF + 
    M.F_M)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.59421 -0.16110 -0.06408  0.16087  0.80493 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.828e+00  1.442e+00   2.654  0.00831 ** 
EDUC        -1.658e-02  6.499e-03  -2.552  0.01114 *  
Age         -6.834e-03  2.075e-03  -3.294  0.00109 ** 
SES         -3.440e-02  1.623e-02  -2.120  0.03466 *  
MMSE        -6.148e-02  4.387e-03 -14.015  < 2e-16 ***
eTIV         2.773e-05  4.814e-04   0.058  0.95410    
nWBV        -1.966e+00  4.540e-01  -4.329 1.95e-05 ***
ASF          2.877e-01  6.053e-01   0.475  0.63484    
M.F_M        9.161e-02  3.165e-02   2.895  0.00403 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2353 on 356 degrees of freedom
Multiple R-squared:  0.5315,	Adjusted R-squared:  0.521 
F-statistic: 50.49 on 8 and 356 DF,  p-value: < 2.2e-16
```

5.3 R Output - Step-Down



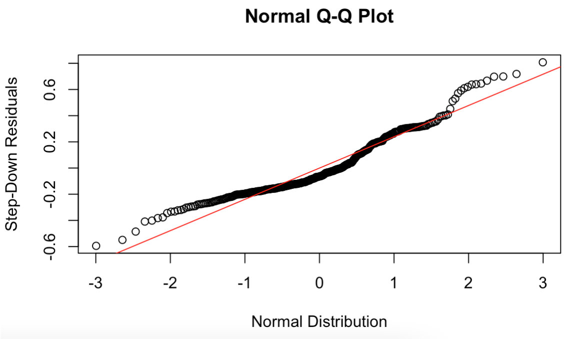
```
Call:
lm(formula = CDR ~ ., data = data[, c("CDR", predictors)])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.59452 -0.16133 -0.06396  0.16125  0.80761 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.907186   0.447822   8.725  < 2e-16 ***
EDUC        -0.016535   0.006435  -2.570  0.01058 *  
Age         -0.006820   0.002057  -3.315  0.00101 ** 
SES         -0.034384   0.016199  -2.123  0.03448 *  
MMSE        -0.061478   0.004380 -14.036  < 2e-16 ***
nWBV        -1.964077   0.452549  -4.340 1.86e-05 ***
ASF          0.253444   0.111603   2.271  0.02375 *  
M.F_M        0.091891   0.031233   2.942  0.00347 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.235 on 357 degrees of freedom
Multiple R-squared:  0.5315,	Adjusted R-squared:  0.5223 
F-statistic: 57.86 on 7 and 357 DF,  p-value: < 2.2e-16
```

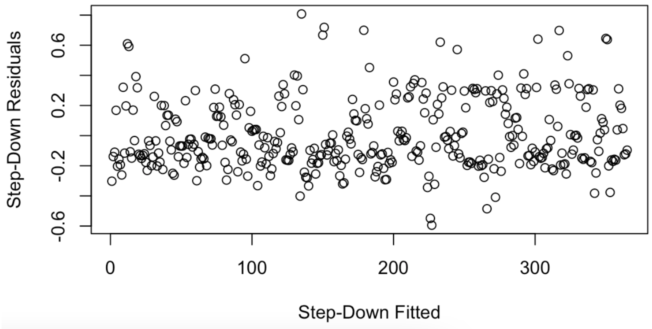
5.4 R Output - LASSO

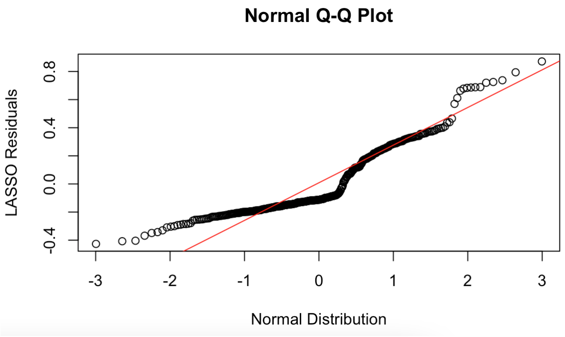


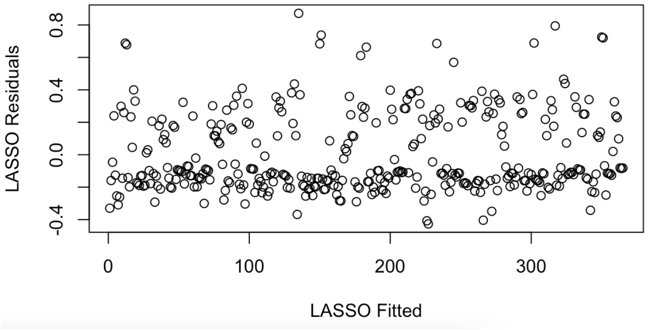
```
9 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept)  3.854779e+00
Age         -6.548741e-03
EDUC        -1.538982e-02
SES         -3.119116e-02
MMSE        -6.148661e-02
eTIV        -1.523943e-06
nWBV        -1.918930e+00
ASF          2.351918e-01
M.F_M        8.809269e-02
```

5.5 Checking Normality and Homoscedasticity of Residuals

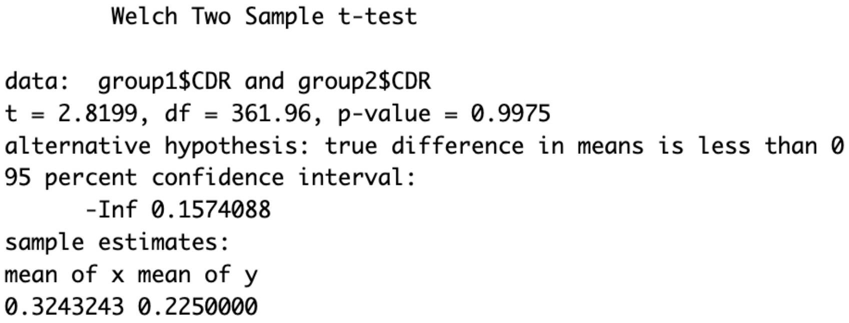








5.6 Hypothesis test



```
        Welch Two Sample t-test

data:  group1$CDR and group2$CDR
t = 2.8199, df = 361.96, p-value = 0.9975
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.1574088
sample estimates:
mean of x mean of y 
0.3243243 0.2250000 
```