What Kills Our Mind? Unveiling the Mysteries with Longitudinal Insights from the Oasis Dataset Study

Eduard Hogea
West University of Timisoara
eduard.hogea00@e-uvt.ro

Abstract

In this research project made for the Statistical Methods for Clinical Studies, I undertake a comprehensive longitudinal analysis of the OASIS MRI/Alzheimer's dataset, an extensive collection of data from 150 subjects aged 60 to 96, including 373 imaging sessions. The dataset encompasses multiple T1-weighted MRI scans for each subject, spanning various visits, and includes variables such as demographics, clinical information, and derived anatomic volumes. The study's core objective is to analyze the progression of dementia, with a particular focus on Alzheimer's disease. Key indicators such as Mini-Mental State Examination (MMSE) scores and Clinical Dementia Rating (CDR) are used to evaluate cognitive impairment and dementia severity. The analysis made in such a way to explore correlations between dementia progression and factors like age, education, socioeconomic status, and brain volume metrics. A pivotal aspect of this research is the comparison between two statistical methodologies: the Random Coefficients Model and the Mean Response Model. Employing Generalized Linear Mixed Models (GLMM) and Ordinal Logistic Regression, these models are chosen for their aptness in handling the longitudinal and categorical nature of the data. The research methodology includes careful handling of missing values, scaling of continuous predictors, and computation of correlation matrices. This study aims to elucidate patterns in dementia progression, with a focus on the efficacy of each model in interpreting the dynamic changes over time in a longitudinal dataset. The results section presents detailed findings, highlighting the comparative performance of the two models in capturing the complexities of Alzheimer's disease progression. The Random Coefficients model outperforms the Mean Response one, considering the evaluation metrics used.

Keywords Alzheimer's Disease, OASIS MRI Dataset, Longitudinal Analysis, Random Coefficients Model, Mean Response Model, Cognitive Impairment

1 Introduction

This project, serving as the final one for the 'Statistical Methods for Clinical Studies' course, dives into a pressing question in the medical world: What leads to the decline of our cognitive abilities, particularly in Alzheimer's disease? Initially, my plan was to explore the Framingham Heart Study, focusing on heart health. That dataset seemed promising, as it is one of the longest ongoing longitudinal studies under the U.S. Public Health Service. However, due to challenges in accessing the latest data (and the fact that it now requires an approval from the IRB/Ethics Committee from the National Heart, Lung, and Blood Institute), I shifted my focus to another significant area in medical research - the study of Alzheimer's disease using the OASIS MRI/Alzheimer's dataset Marcus et al. [2010]. The latter one is an open source dataset.

Alzheimer's disease, known for causing memory loss and cognitive decline, affects millions worldwide. The OASIS dataset, which includes brain scans and health data of older adults, provides an excellent opportunity to study this disease over time give the nature of the longitudinal approach in data collection. In this project, I aim to understand how Alzheimer's disease progresses, using statistical tools to analyze changes in the brain and cognition of individuals in the dataset.

To do this, I am using two types of longitudinal analysis models: the Random Coefficients Model and the Mean Response Model. The code will also be available with this report. These models help analyze data that changes over time, like the repeated brain scans and health checks in the OASIS dataset. By comparing these models, I hope to find out which one gives us a clearer picture of how Alzheimer's disease progresses.

2 Literature Review

The exploration of Alzheimer's disease using the OASIS dataset has led to significant insights. A popular paper Baglat et al. [2020] developed as a joint research project between universities from India, Portugal and Switzerland demonstrated the effectiveness of Random Forest and Adaptive boosting in early detection of the disease. This study concludes with an 86% accuracy rate, but states that a combination of more complex techniques such as CNNs and SVMs would surely outperform.

In contrast, the survey form this paper Sh et al. [2022] delved into the role of MRI in Alzheimer's, emphasizing structural brain differences, an aspect crucial for early detection and differentiation from healthy brains. This focus on imaging techniques complements the machine learning approaches seen in other studies and also recognizes that Deep Learning models are a necessity in advancing this field, given their achieved high accuracies in literature

Quite interestingly, Random Forest is used once again in this paper Jahan et al. [2023] for Alzheimer's prediction. Their study, achieved a perfect accuracy score, of 100% on the same dataset discussed in this project.

Collectively, these studies present a unified narrative: the MRIs (and datasets containing data from those, such as the OASIS longitudinal dataset) are vital for developing Alzheimer's detection tools, with machine learning, especially Random Forest, playing a key role. While each study has its unique approach, it is agreed that constant monitoring of the disease is a key aspect in finding and slowing the development of it, highlighting also the importance of the longitudinal analysis.

3 Methodology

In this study, longitudinal analysis methods are employed to understand the progression of Alzheimer's disease using the OASIS dataset, and to understand more in depth what features are defining for the evolution of the disease. This involves a detailed investigation of changes in brain structure and cognitive function over time. The analysis hinges on two statistical models: the Random Coefficients Model and the Mean Response Model. The Random Coefficients Model is particularly useful for data that vary across individuals, allowing for the estimation of individual trajectories in cognitive decline. In contrast, the Mean Response Model provides insights into the average trend of the entire population, offering a broader perspective on the disease progression.

4 Dataset Format Description

The dataset comprises a longitudinal study of 150 individuals aged between 60 and 96 years. Each participant underwent MRI scans during two or more visits, with each visit spaced at least a year apart, amounting to a total of 373 imaging sessions. In each session, 3 to 4 T1-weighted MRI scans were performed on each subject. All participants were right-handed and the group was composed of both males and females. Throughout the course of the study, 72 subjects consistently showed no signs of dementia. However, 64 subjects were identified with dementia from their first visit, including 51 with mild to moderate Alzheimer's disease. Additionally, 14 subjects initially assessed as nondemented were later diagnosed with dementia in subsequent visits. The data was curated by dropping the rows with any missing values. The summary of the data can be seen in Table 1.

5 Data Analysis

In the analysis of the OASIS longitudinal dataset, specific data features were carefully selected for their relevance to Alzheimer's disease progression and diagnosis, while certain features were purposefully excluded. The chosen features included MMSE (Mini-Mental State Examination) scores, CDR (Clinical Dementia Rating), age, educational level (EDUC), socioeconomic status (SES), normalized whole-brain volume (nWBV), and estimated total intracranial volume (eTIV). These variables were chosen to explore the cognitive impairment, dementia severity, age-related risk, socio-demographic patterns, and neuroimaging measurements in relation to Alzheimer's disease. Contrary, features such as handedness, gender, MRI ID, subject ID, and MR Delay were omitted from the primary analysis due to their limited relevance to the study's specific research objectives. This careful selection of features and exclusion of irrelevant

Table 1: Summary of the OASIS Longitudinal Dataset

Variable	**Description**	**Value Range or Unique Values**
Subject.ID	Unique identifier for each subject	150 unique IDs
MRI.ID	Identifier for each MRI scan	373 unique IDs
Group	Classification of subject	Nondemented, Demented, Converted
Visit	Number indicating the visit sequence	1 to 5
MR.Delay	Time delay between MRI scans	0 to 2639 days
M.F	Gender of the subject	Male (M), Female (F)
Hand	Handedness (all right-handed)	Right (R)
Age	Age of the subject	60 to 98 years
EDUC	Years of education	6 to 23 years
SES	Socioeconomic status	1 (highest) to 5 (lowest)
MMSE	Mini-Mental State Examination score	0 to 30
CDR	Clinical Dementia Rating	0 to 2
eTIV	Estimated total intracranial volume in mm³	1106 to 2004 mm³
nWBV	Normalized whole-brain volume	64.4% to 83.7%
ASF	Atlas scaling factor	0.876 to 1.587

ones aimed to uncover nuanced relationships within the dataset, contributing to a deeper understanding of Alzheimer's disease progression and diagnosis. More in depth analysis for these features can be found in Figures 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.

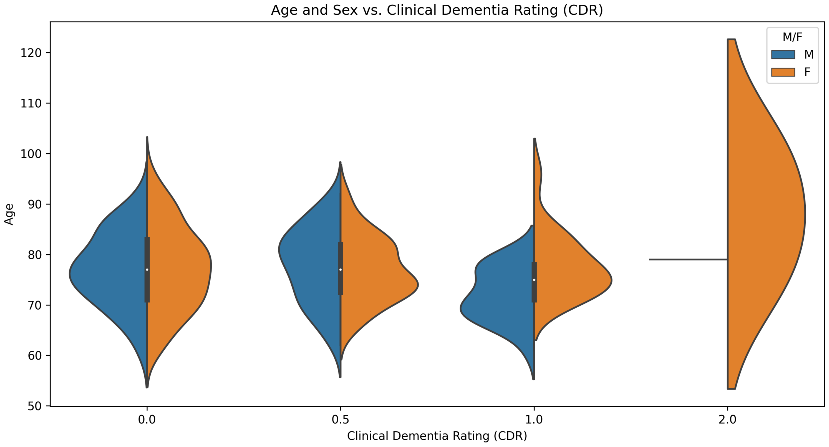

Figure 1: Age and Sex vs. Clinical Dementia Rating (CDR): This plot illustrates the distribution of age with respect to the clinical dementia rating, differentiated by sex. Observation: There is no obvious connection between Age/Sex and Dementia Diagnosis(if we ignore that for 2.0 CDR, ther are no male samples.

Some interesting findings from the correlation plot in Figure 11.

- There is a strong negative correlation (-0.526) between Age and nWBV (normalized whole brain volume), indicating that as individuals age, their whole brain volume tends to decrease.
- EDUC (years of education) and SES (socioeconomic status) show a strong negative correlation (-0.722), suggesting that individuals with higher years of education tend to have a lower socioeconomic status.
- MMSE (Mini-Mental State Examination) has a moderate positive correlation (0.341) with nWBV, indicating that individuals with higher normalized whole brain volumes tend to score better on the MMSE, a measure of cognitive function.
- SES has a negative correlation (-0.261) with eTIV (estimated total intracranial volume), implying that individuals with lower socioeconomic status tend to have smaller estimated total intracranial volumes.

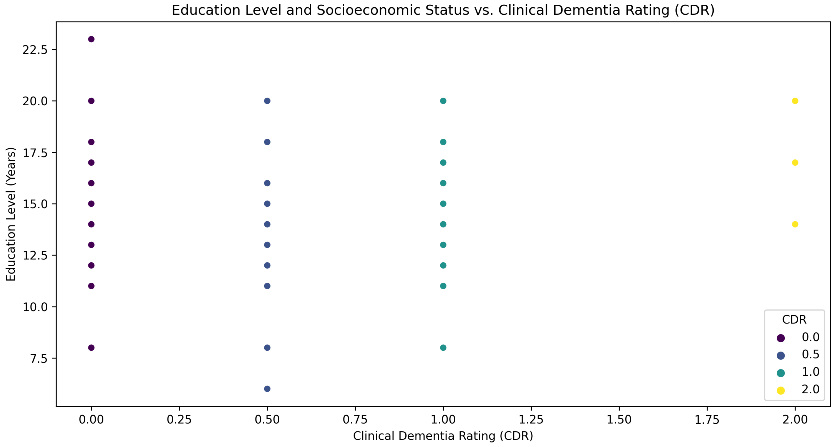

Figure 2: Education Level and Socioeconomic Status vs. Clinical Dementia Rating (CDR): This scatter plot shows the relationship between education level and clinical dementia rating, with each point representing a participant. Observation: There is still no obvious connection between Education Level/Social Economic Status and Dementia Diagnosis.

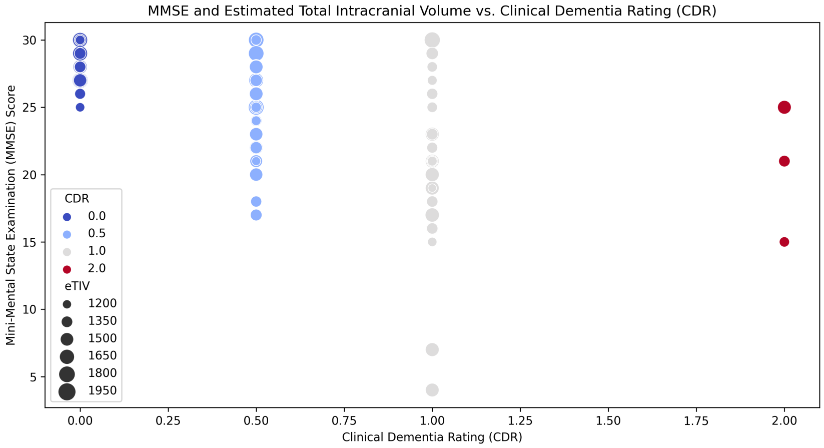

Figure 3: MMSE and Estimated Total Intracranial Volume vs. Clinical Dementia Rating (CDR): This plot visualizes the MMSE scores of participants against their clinical dementia rating, with the size of each point representing the estimated total intracranial volume. Observation: While the MMSE examination results of subjects not diagnosed with dementia concentrate near 27-30 points, results for diagnosed subjects are more spread out, including cases with high MMSE scores but a CDR of 0.5 or 1. There is no obvious connection between Estimated Total Intracranial Volume and Dementia Diagnosis.

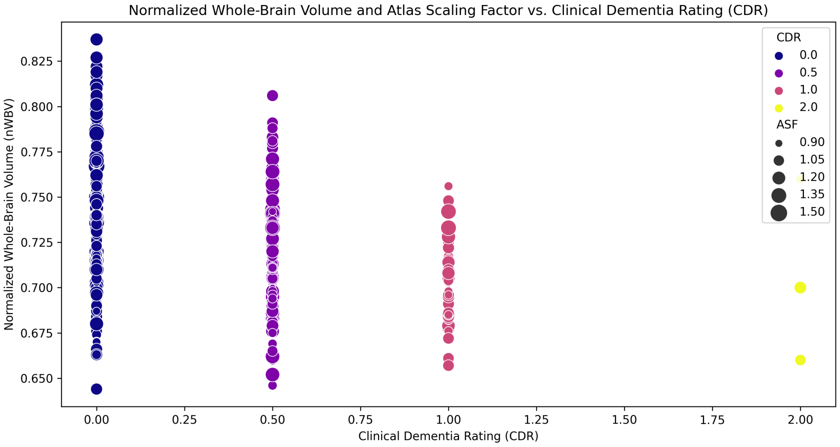

Figure 4: Normalized Whole-Brain Volume and Atlas Scaling Factor vs. Clinical Dementia Rating (CDR): This scatter plot illustrates the relationship between normalized whole-brain volume and clinical dementia rating, with the size of each point indicating the atlas scaling factor. Observation: Normalized whole-brain volume appears more spread out for subjects with a CDR of 0 and becomes narrower as CDR increases. There is no obvious connection between Atlas Scaling Factor and Dementia Diagnosis.

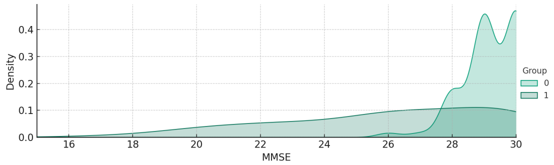

Figure 5: MMSE Distribution. The chart shows that the Nondemented group has higher MMSE scores than the Demented group.

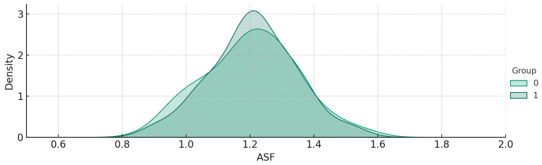

Figure 6: ASF Distribution.

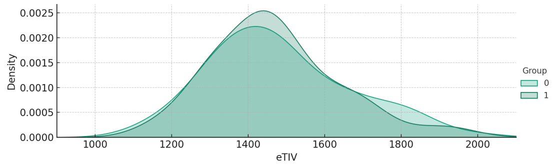

Figure 7: eTIV Distribution.

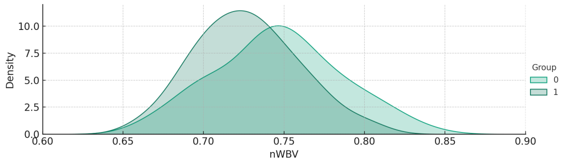

Figure 8: nWBV Distribution. The chart indicates that the Nondemented group has a higher brain volume ratio than the Demented group.

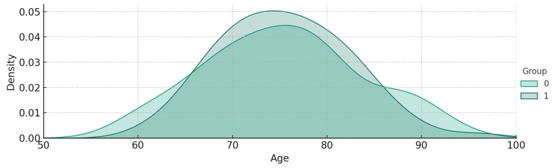

Figure 9: Age Distribution. There is a higher concentration of 70-80 years old in the Demented patient group than those in the nondemented patients.

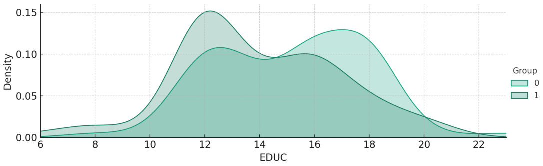

Figure 10: EDUC Distribution. Demented patients were less educated in terms of years of education.

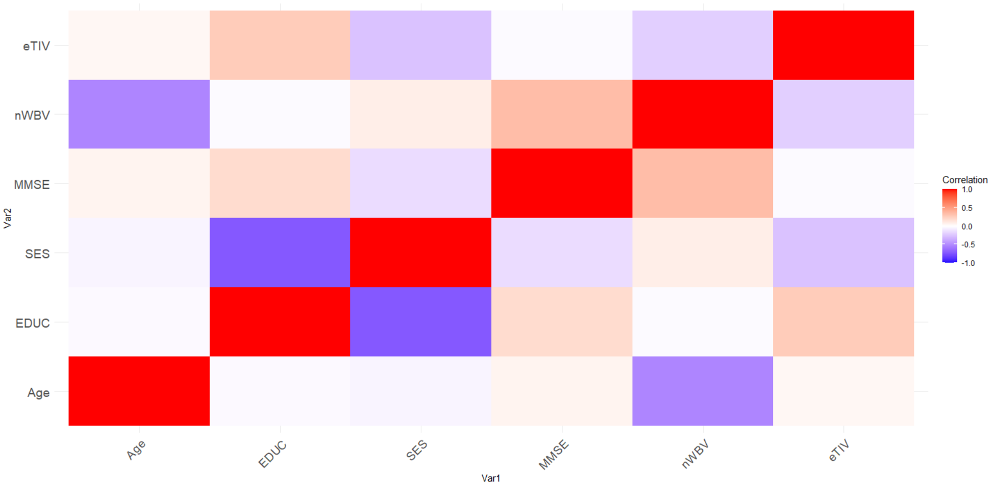

Figure 11: Correlation plot of the remaining relevant features.

(a) Age 60-69

Metric	Value
MMSE Scores	
Mean	26.88
Std Dev	5.43
Range	4 - 30
nWBV	
Mean	0.761
Std Dev	0.041
Range	0.676 - 0.837
CDR	
Mean	0.27
Std Dev	0.37
Range	0 - 1

(b) Age 70-79

Metric	Value
MMSE Scores	
Mean	27.28
Std Dev	3.44
Range	15 - 30
nWBV	
Mean	0.734
Std Dev	0.031
Range	0.657 - 0.810
CDR	
Mean	0.33
Std Dev	0.41
Range	0 - 2

(c) Age 80-89

Metric	Value
MMSE Scores	
Mean	27.66
Std Dev	2.91
Range	17 - 30
nWBV	
Mean	0.712
Std Dev	0.029
Range	0.644 - 0.762
CDR	
Mean	0.25
Std Dev	0.30
Range	0 - 1

Figure 12: **Age 60-69:** In this age group, there is variability in MMSE scores, indicating differences in cognitive performance among individuals. The nWBV values reflect changes in brain volume, with individual variations. CDR scores suggest the presence of dementia symptoms, but the severity varies.
Age 70-79: Within this age range, we observe greater variability in MMSE scores, suggesting increased diversity in cognitive scores, including some decline compared to the previous group. nWBV measurements continue to show changes in brain volume, emphasizing individual differences. CDR scores reveal a wider range of dementia symptoms, with some individuals experiencing more significant progression.
Age 80-89: In the oldest age group, MMSE scores tend to exhibit more pronounced changes, indicating a greater impact on cognitive abilities. Additionally, the decline in brain volume, as indicated by nWBV, becomes more noticeable, potentially reflecting age-related structural changes in the brain. CDR scores demonstrate both increased severity and variability in dementia symptoms, underscoring the complexity of the condition among individuals in this advanced age group.

6 Experimentation results

The analysis and the experiments focused on the remaining significant variables. Preprocessing included handling missing values, scaling continuous predictors, and categorizing CDR. For the evaluation, I have used the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to compare the models. These criteria are defined as follows:

AIC (Akaike Information Criterion)

- **Purpose:** AIC is used to compare different potential models, determining which best explains the data while penalizing the number of parameters.
- **Calculation:** AIC is calculated using the formula AIC $= 2k - 2\ln(L)$, where k is the number of parameters in the model and L is the likelihood of the model. A lower AIC value indicates a better model.
- **Usage:** AIC is particularly useful in assessing the prediction accuracy of the model, balancing between model complexity and goodness of fit.

BIC (Bayesian Information Criterion)

- **Purpose:** BIC is used to compare models but includes a stronger penalty for models with more parameters, guarding against overfitting.
- **Calculation:** BIC is calculated with BIC $= \ln(n)k - 2\ln(L)$, where n is the number of observations, k the number of parameters, and L the likelihood of the model. A lower BIC value suggests a better model.
- **Usage:** BIC is often used in Bayesian model selection and is more appropriate for larger datasets, where overfitting is a concern.

The Key Differences are that BIC penalizes model complexity more heavily than AIC, especially in large samples and AIC focuses on identifying the model that best predicts the data, while BIC is geared towards finding the true model. The Random Coefficients Model was developed within a Generalized Linear Mixed Model (GLMM) framework and the Mean Response Model using ordinal logistic regression. The Random Coefficients Model, represented as:

$$\text{CDR} \sim \text{MMSE} + \text{nWBV} + \text{Age} + \text{Visit} + (1|\text{SubjectID})$$

enabled the understanding of both fixed effects and random individual variations in dementia progression. This model showed significant relationships between the predictors and CDR, with MMSE, nWBV, and Age displaying negative coefficients, and Visit showing a positive relationship. The model also suggested substantial individual variability, as indicated by the random effects analysis.

The Mean Response Model, formulated as:

$$\text{CDR} \sim \text{MMSE} + \text{nWBV} + \text{Age} + \text{Visit}$$

provided insights into the average impact of these predictors on CDR, without considering individual differences. This model revealed a similar trend in the significance and direction of predictors, offering a general overview of how each factor is associated with dementia progression. Both models were robust in their explanatory power, with the Random Coefficients Model achieving an AIC of 372.69 and a BIC of 403.64, while the Mean Response Model recorded an AIC of 458.94 and a BIC of 486.03. These values indicate a good balance between model complexity and fit.

The findings from each model were as follows:

- Random Coefficients Model: Significant relationships were observed between CDR and all predictors. MMSE, nWBV, and Age showed negative coefficients, indicating their inverse relationship with CDR. Visit, however, had a positive relationship with CDR.
- Mean Response Model: This model highlighted the average impact of predictors on CDR, revealing similar trends in the significance and direction of coefficients as observed in the Random Coefficients Model.

Additional analysis for subjects in the 'Converted' group was conducted. The idea was to focus only on the persons that suffered a fall into dementia, the logic here being that there must have been something wrong with them even before attending the visit where the disease was discovered. The AIC and BIC for both models in this group showed notably lower values, indicating a different progression pattern within this subset. The models also seem to fit the data better when only focusing on the converted cases, and the final table with the results of those models is in Table 2.

An unfortunate observed trend I found is that of an decreasing MMSE value for persons that already have Alzheimer. This can be seen in Figure 13, where all the patients with 3 or more visits have a decrease in their final MMSE as opposed to its value at the first visit.

Model	AIC	BIC
Random Coefficients Model	372.69	403.64
Mean Response Model	458.94	486.03
RC Model (Converted Group)	36.39	46.05
MR Model (Converted Group)	34.39	42.44
RC Model (Converted Progression)	36.39	46.05
MR Model (Converted Progression)	34.39	42.44

Table 2: AIC and BIC Results

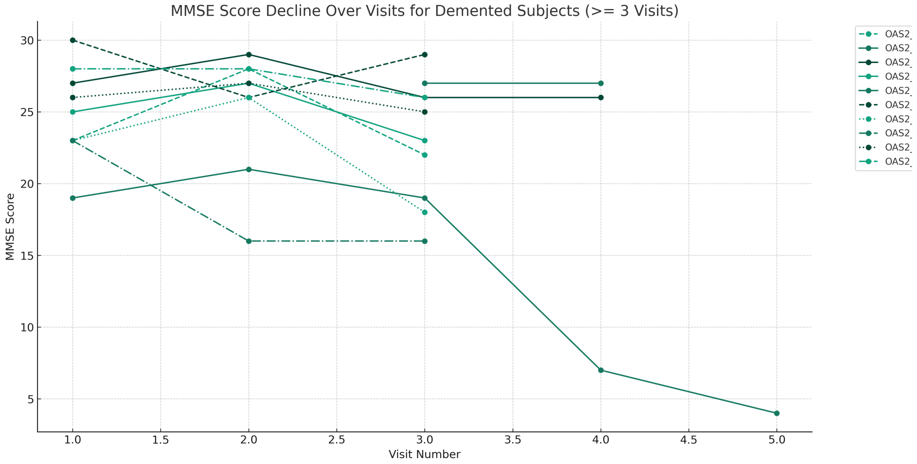

Figure 13: Evolution of MMSE in Demented Subjects. All the subjects ended up with a lower MMSE value that the one they have started with, in the cases of equal or more than 3 visits.

7 Conclusion

In this study, I have successfully applied statistical models to the OASIS dataset to analyze the progression of Alzheimer's disease. The analysis highlighted several key findings:

- The Random Coefficients Model, which accounts for individual variability, suggested a significant negative relationship between Alzheimer's disease progression (as measured by the Clinical Dementia Rating) and factors like MMSE scores, normalized whole brain volume (nWBV), and age. The positive relationship with the visit number indicated a gradual increase in the severity of dementia over time.
- The Mean Response Model provided a broader perspective, indicating general trends across the entire dataset. Similar to the Random Coefficients Model, it pointed out the negative correlation of MMSE, nWBV, and age with dementia progression.
- A focused analysis on the 'Converted' group revealed a different pattern of disease progression. The models showed better fit for this subset, suggesting that specific factors might influence the progression of Alzheimer's disease in individuals initially classified as nondemented.
- The longitudinal analysis, especially the examination of MMSE scores over time in subjects with dementia, underscored the progressive nature of cognitive decline in Alzheimer's disease. This was particularly evident in the subset of patients with three or more visits, where a consistent decrease in MMSE scores was observed.

These findings helped me have a better understanding of Alzheimer's disease, and also showcased a hands-on appliance of the theoretical parts discussed at the lectures regarding the longitudinal studies in capturing the nuances of such complex medical conditions.

However, I don't consider this dataset to be one of the highest quality. There are multiples instances of samples missing important features or patients that have started from 2nd or 3rd visit, essentially making the dataset incomplete. The fact that one of the cited papers in the literature overview also showed 100% accuracy with a relatively simple model for classification also makes me doubt that a medical problem such as this one is that easily solvable with a Random Forest model.

In conclusion, this project has been interesting and the appliance of longitudinal analysis seemed to be a powerful technique that is quite easily implemented in R, but for more refined plots I have moved back to Python. Thank you for teaching us the course of Statistical Methods for Clinical Studies!

References

Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12):2677–2684, 2010.

P. Baglat et al. Multiple machine learning models for detection of alzheimer's disease using oasis dataset. *Journal of Medical Systems*, 44(2):37, 2020. doi:10.1007/s10916-018-1088-1.

Zahraa Sh, Aaraji, Hawraa Abbas, and Ameer Asady. Alzheimer's diseases detection by using mri brain images: A survey. 2022.

Sobhana Jahan, Kazi Abu Taher, M. Shamim Kaiser, and In-ho Ra. Explainable ai-based alzheimer's prediction and management using multimodal data. *PLOS ONE*, 2023.