Project 1: Contrasting Aging Factors with Alzheimer's Development

Lavanyaa Gupta

2024-10-22

Background

What is Alzheimer's Disease?

- According to the Alzheimer's Association, Alzheimer's is a "type of dementia that affects memory, thinking, and behavior"
- There is currently no cure for Alzheimer's, however, understanding the affliction and recognizing progression can allow for early interventions.
- Therefore, data is being collected to understand differences in young and aged Alzheimer's patients.

Data Set Description

- Subject ID: unique participant identifier
- M/F: identifies male or female participant
- Hand: right/left hand dominance. Typically, scientific studies do not include left hand dominant participants.
- Educ: refers to education level; the manner in which numbers are attributed is unknown.
- SES: socioeconomic status
- MMSE: mini-mental state examination
- CDR: clinical dementia rating
- eTIV: estimated total intracranial volume
- nWBV: Normalize Whole Brain Volume
- ASF: Atlas Scaling Factor
- Delay: unknown, likely some type of error

Hypothesis

- Higher SES and Educ are associated with lower CDR scores, indicating slower progression of Alzheimer's disease, even after adjusting for brain volume measures eTIV, nWBV and MMSE.
- The problem here is that Alzheimer's is rapidly progressing and recognizing patterns and warning signs is vital.

Set Up

```r
library(readr)
oasis_cross_sectional <-read.csv("C:/Users/lavan/Downloads/oasis_cross-sectional.csv")
View(oasis_cross_sectional)
```

- We have now loaded in the CROSS-SECTIONAL data. We will examine the LONGITUDINAL data later.
- Now, let's import libraries.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v purrr     1.0.2
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(ggplot2)
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.4.1

## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```r
library(lmerTest)
```

```
## Warning: package 'lmerTest' was built under R version 4.4.1

##
## Attaching package: 'lmerTest'
##
## The following object is masked from 'package:lme4':
##
##     lmer
##
## The following object is masked from 'package:stats':
##
##     step
```

Visualizing the Data & Preliminary Analysis

Basic Data Wrangling to Ensure Data Set is Ready for Models and Plotting

```r
oasis_cross_sectional <- oasis_cross_sectional %>%
  na.omit()
```

Analysis of Cross-Sectional Data to Understand Trends and Distribution

```r
SES_mean <- mean(oasis_cross_sectional$SES)
SES_sd <- sd(oasis_cross_sectional$SES)

Educ_mean <- mean(oasis_cross_sectional$Educ)
Educ_sd <- sd(oasis_cross_sectional$Educ)

oasis_cross_sectional <- oasis_cross_sectional %>%
  filter(SES > (SES_mean - 3 * SES_sd) & SES < (SES_mean + 3 * SES_sd)) %>%
  filter(Educ > (Educ_mean - 3 * Educ_sd) & Educ < (Educ_mean + 3 * Educ_sd))

oasis_cross_sectional <- oasis_cross_sectional %>%
  mutate(Age_Group = cut(Age, breaks = c(50, 60, 70, 80, 90, 100), right = FALSE, labels = c("50-59", "6
```

Exploratory Statistics on Cross-Sectional Data

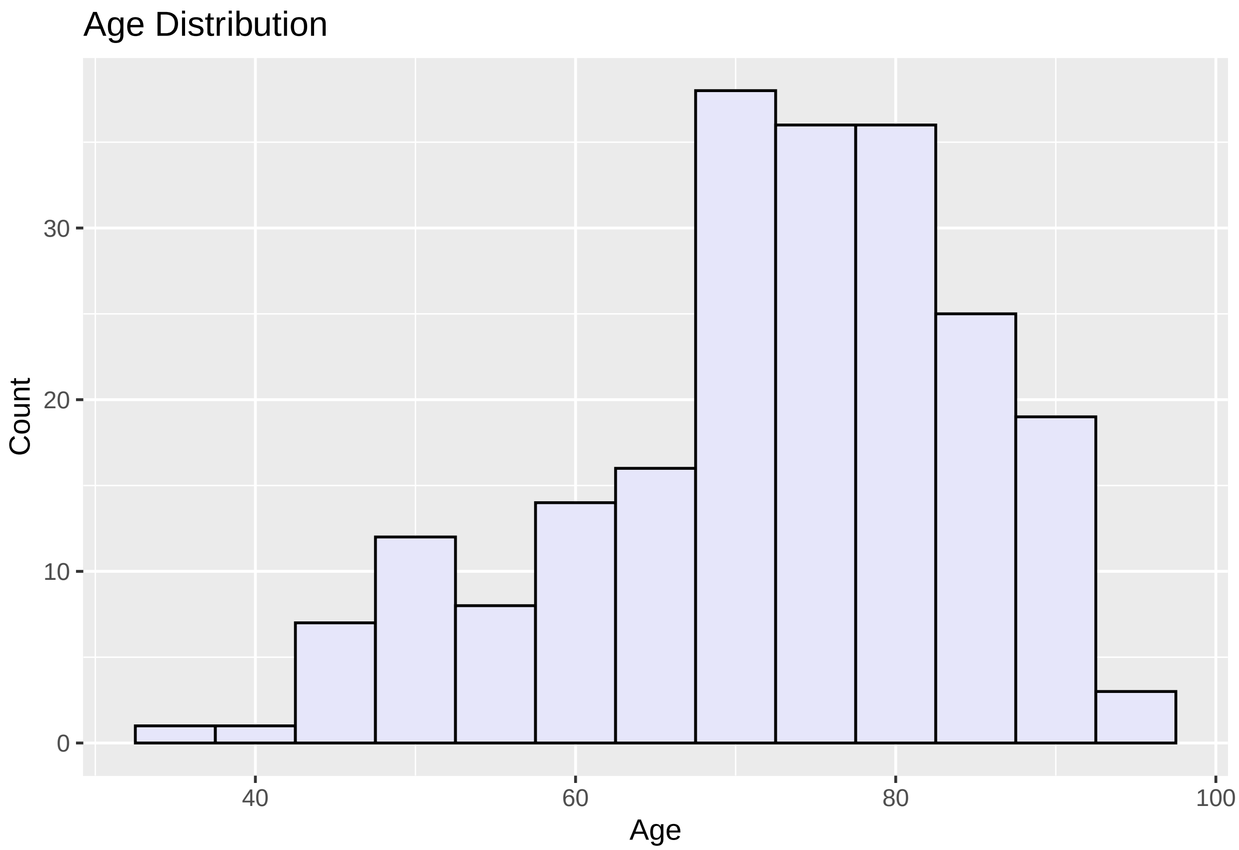
```r
summary_stats <- oasis_cross_sectional %>%
  summarise(
    Age_mean = mean(Age),
    Age_sd = sd(Age),
    MMSE_mean = mean(MMSE),
    MMSE_sd = sd(MMSE),
    nWBV_mean = mean(nWBV),
    nWBV_sd = sd(nWBV)
  )

summary_stats
```

```
##   Age_mean   Age_sd MMSE_mean  MMSE_sd nWBV_mean    nWBV_sd
## 1 72.44444 12.30642  27.32407 3.43668    0.7505 0.04827104
```

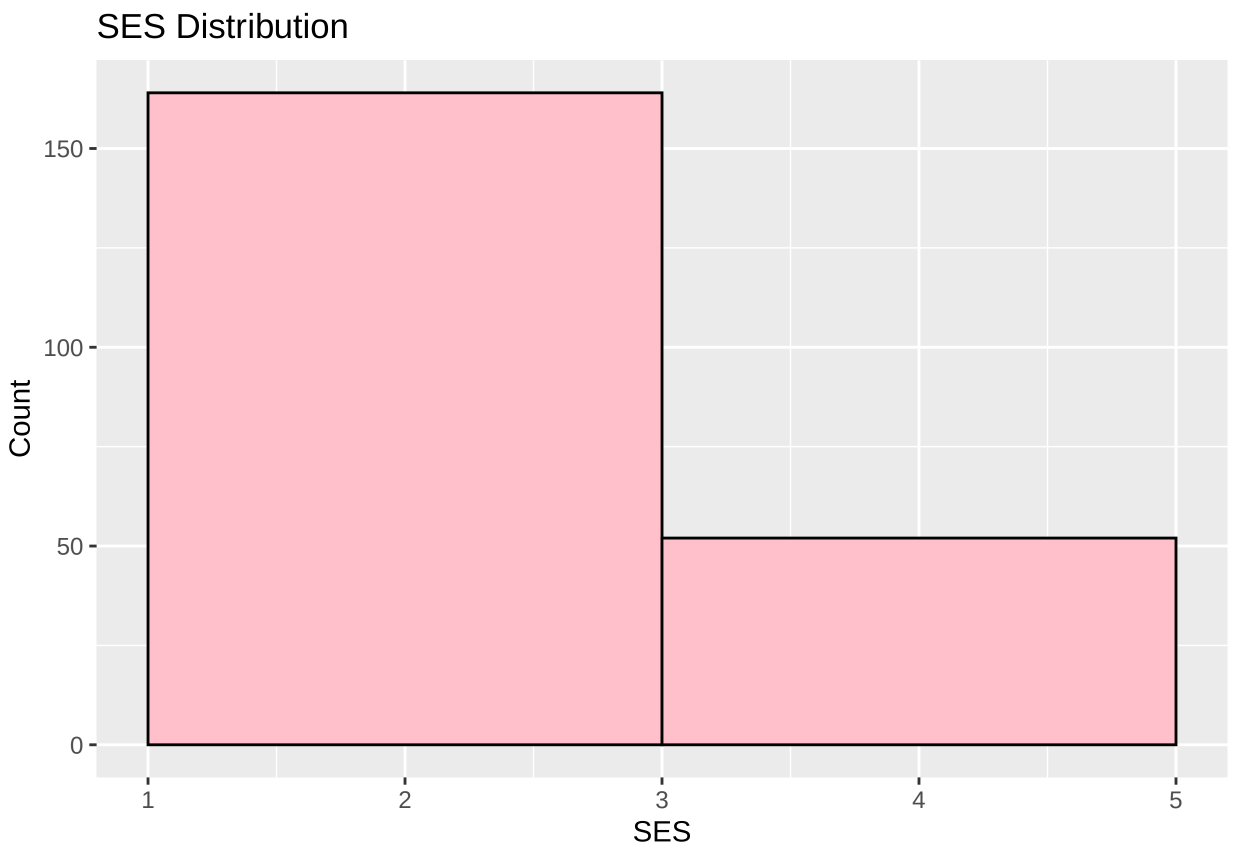
Ditribution Analysis

```r
distrib_Age <- ggplot(oasis_cross_sectional, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lavender", color = "black") +
  labs(title = "Age Distribution", x = "Age", y = "Count")

distrib_SES <- ggplot(oasis_cross_sectional, aes(x = SES)) +
  geom_histogram(binwidth = 2, fill = "pink", color = "black") +
  labs(title = "SES Distribution", x = "SES", y = "Count")

distrib_EDUC <- ggplot(oasis_cross_sectional, aes(x = Educ)) +
  geom_histogram(binwidth = 0.01, fill = "purple", color = "black") +
  labs(title = "EDUC Distribution", x = "EDUC", y = "Count")
```

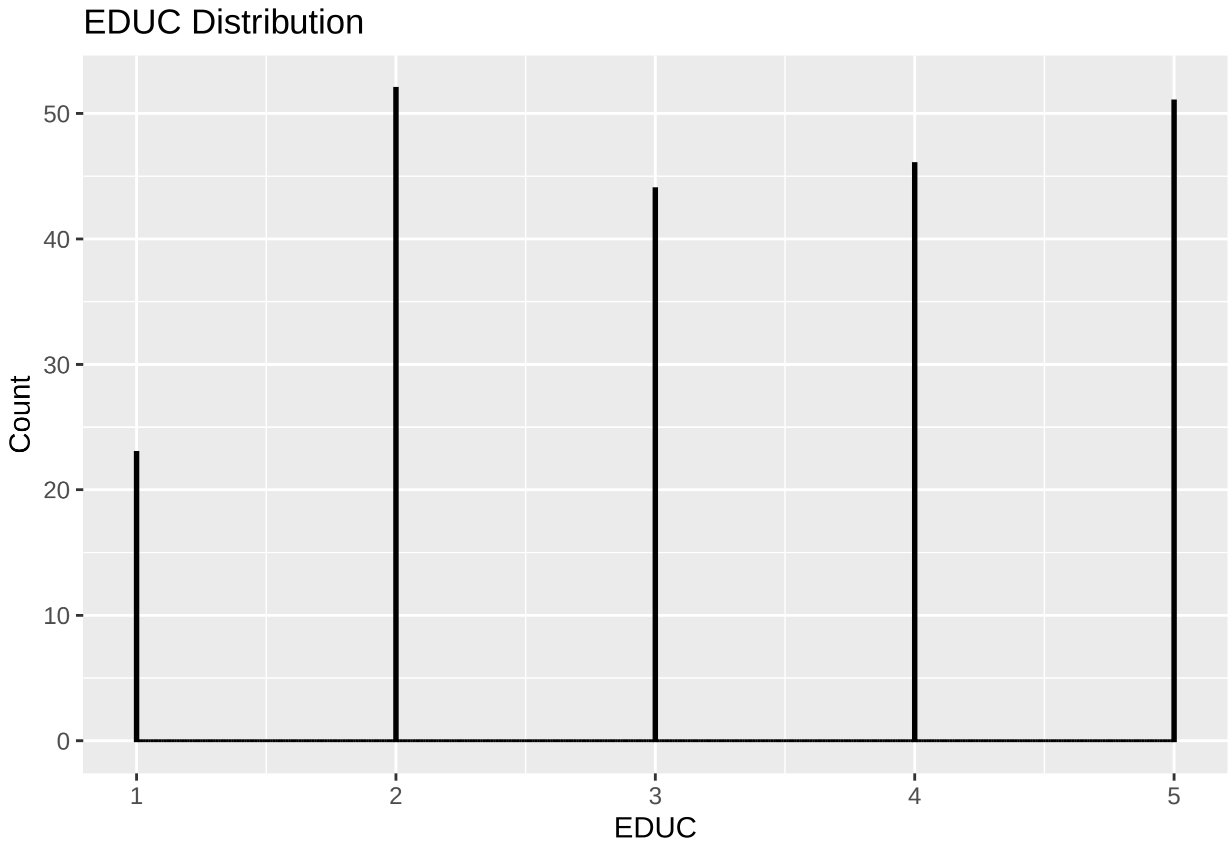
```
distrib_Age
```



```
distrib_SES
```



```
distrib_EDUC
```



Building a Linear Model to Predict CDR Scores as Correlated by CDR and SES

```
model_CDR <- lm(CDR ~ SES + Educ + MMSE + eTIV + nWBV, data = oasis_cross_sectional)
summary(model_CDR)
```

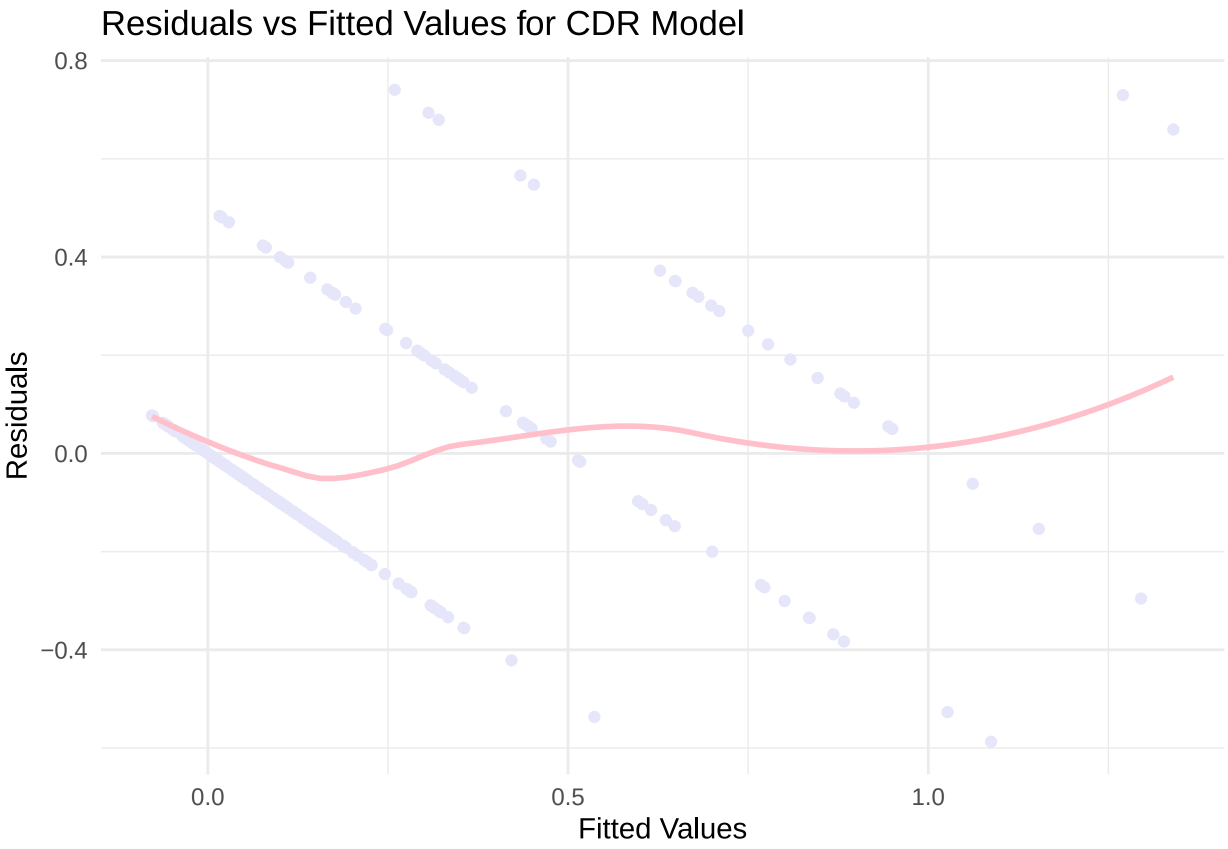
```
##
## Call:
## lm(formula = CDR ~ SES + Educ + MMSE + eTIV + nWBV, data = oasis_cross_sectional)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58720 -0.14230 -0.03344  0.07962  0.74057
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.9691200  0.3462599   8.575 2.18e-15 ***
## SES          0.0037244  0.0215852   0.173  0.86317
## Educ        -0.0012046  0.0183380  -0.066  0.94769
## MMSE        -0.0781779  0.0055384 -14.116  < 2e-16 ***
## eTIV         0.0001873  0.0001056   1.774  0.07758 .
## nWBV        -1.1326298  0.3929567  -2.882  0.00436 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.236 on 210 degrees of freedom
## Multiple R-squared:  0.6282, Adjusted R-squared:  0.6194
## F-statistic: 70.96 on 5 and 210 DF,  p-value: < 2.2e-16
```

Ploting the Linear Model

```
fitted_values <- fitted(model_CDR)
residuals <- resid(model_CDR)

plot1<- ggplot(data = NULL, aes(x = fitted_values, y = residuals)) +
  geom_point(color = "lavender") +
  geom_smooth(method = "loess", color = "pink", se = FALSE) +
  labs(title = "Residuals vs Fitted Values for CDR Model",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()

plot1
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Building an ANOVA (Analysis of Variance) Model to Account for Age

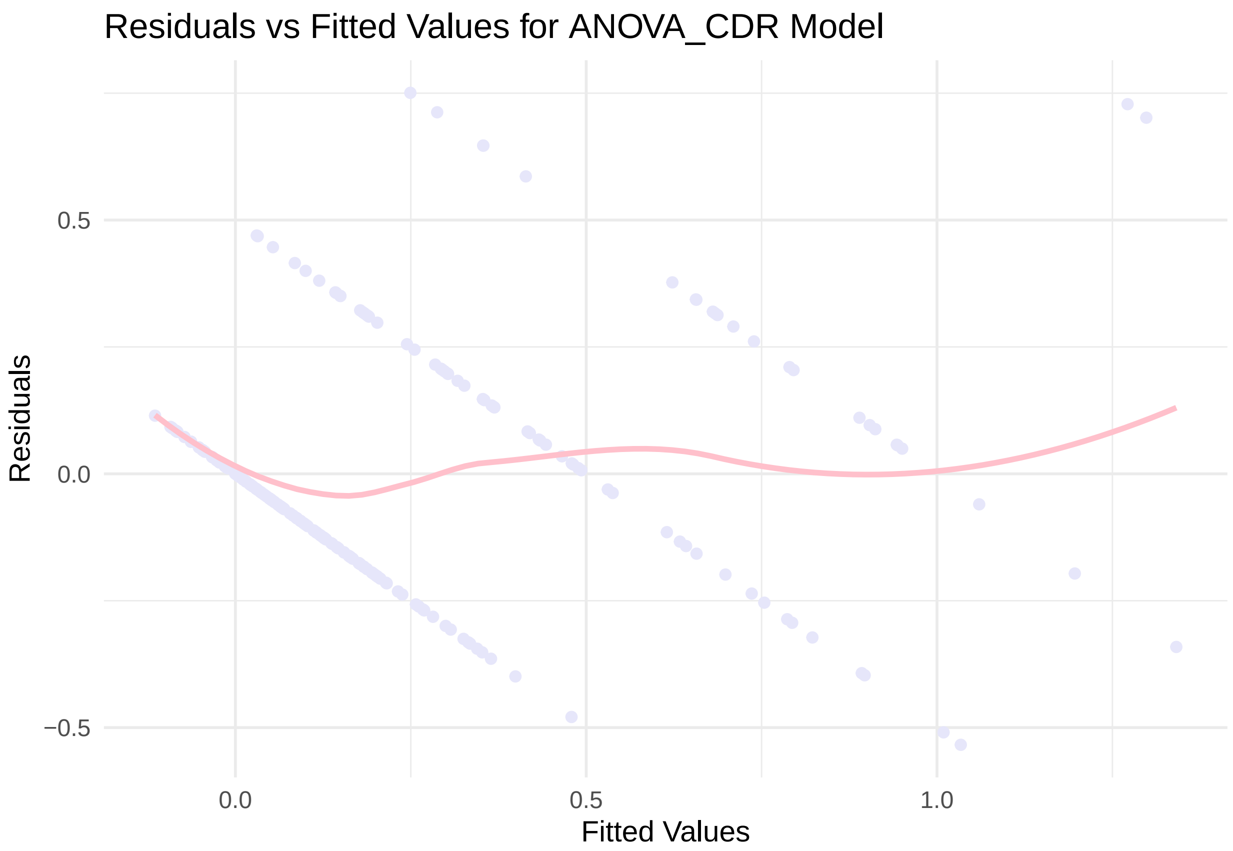
```
ANOVA_CDR <- lm(CDR ~ SES + Educ + MMSE + eTIV + nWBV * Age, data = oasis_cross_sectional)
summary(ANOVA_CDR)
```

```
##
## Call:
## lm(formula = CDR ~ SES + Educ + MMSE + eTIV + nWBV * Age, data = oasis_cross_sectional)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53392 -0.14283 -0.03668  0.08843  0.75065
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.8473934  1.9004997   3.077  0.00237 **
## SES          0.0004358  0.0217090   0.020  0.98400
## Educ        -0.0037980  0.0184009  -0.206  0.83668
## MMSE        -0.0771793  0.0055857 -13.817  < 2e-16 ***
## eTIV         0.0001636  0.0001073   1.524  0.12896
## nWBV        -4.6719768  2.3699821  -1.971  0.05001 .
## Age         -0.0348180  0.0236453  -1.473  0.14239
## nWBV:Age     0.0430953  0.0303047   1.422  0.15650
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2358 on 208 degrees of freedom
## Multiple R-squared:  0.6325, Adjusted R-squared:  0.6202
## F-statistic: 51.15 on 7 and 208 DF,  p-value: < 2.2e-16
```

Plot the ANOVA Model

```
fitted_values <- fitted(ANOVA_CDR)
residuals <- resid(ANOVA_CDR)

plot2 <- ggplot(data = NULL, aes(x = fitted_values, y = residuals)) +
  geom_point(color = "lavender") +
  geom_smooth(method = "loess", color = "pink", se = FALSE) +
  labs(title = "Residuals vs Fitted Values for ANOVA_CDR Model",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()

plot2
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Longitudinal Data Set Analysis

- Along with the cross-sectional Alzheimer's data found on Kaggle, I also gained access to longitudinal data with the same sample set.
- Observing Longitudinal data sets is imperative in progression research (ex: blood cancer progression, Alzheimer's progression) to understand what minor physiological changes may be occurring that could account for the rapid changes
- Or, how SES or Educ could perhaps be involved in progression as well

Data Set Description

- Subject ID: unique participant identifier
- MRI ID: unique MRI scan identifier that corresponds to Subject ID
- Group: Identifies whether patient presents Dementia symptoms
- Visit: this data set is longitudinal, therefore, each patient may attend more than one visit to the particular clinic. Multiple visits allows us to monitor progression signs.
- MR Delay: unknown
- Hand, M/F, Age, EDUC, MMSE, CDR, eTIV, nWB, and ASF are all the same as cross-sectional

Set Up

```
oasis_longitudinal <- read.csv("C:/Users/lavan/Downloads/oasis_longitudinal.csv")
View(oasis_longitudinal)
```

Building a Mixed-Effect Model

```
mixed_model <- lmer(MMSE ~ Group + Age + nWBV + Visit + (1 | Subject_ID), data = oasis_longitudinal)
summary(mixed_model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: MMSE ~ Group + Age + nWBV + Visit + (1 | Subject_ID)
##    Data: oasis_longitudinal
##
## REML criterion at convergence: 1699.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.6626 -0.3012  0.0439  0.3966  3.0754
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  Subject_ID (Intercept) 4.666    2.160
##  Residual               3.111    1.764
## Number of obs: 371, groups:  Subject_ID, 150
##
## Fixed effects:
##                    Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)         4.16393    6.30848 205.19684   0.660  0.50996
## GroupDemented      -3.57368    0.73202 145.23366  -4.882 2.74e-06 ***
## GroupNondemented    0.36842    0.71426 142.23732   0.516  0.60679
## Age                 0.07548    0.03117 166.70955   2.421  0.01654 *
## nWBV               26.36788    6.31027 214.16748   4.179 4.27e-05 ***
## Visit              -0.36782    0.12357 321.72273  -2.977  0.00314 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) GrpDmn GrpNnd Age    nWBV
## GroupDemntd -0.267
## GropNndmntd -0.050  0.804
## Age         -0.772  0.157  0.024
## nWBV        -0.943  0.155 -0.071  0.536
## Visit       -0.060  0.029 -0.030 -0.235  0.157
```

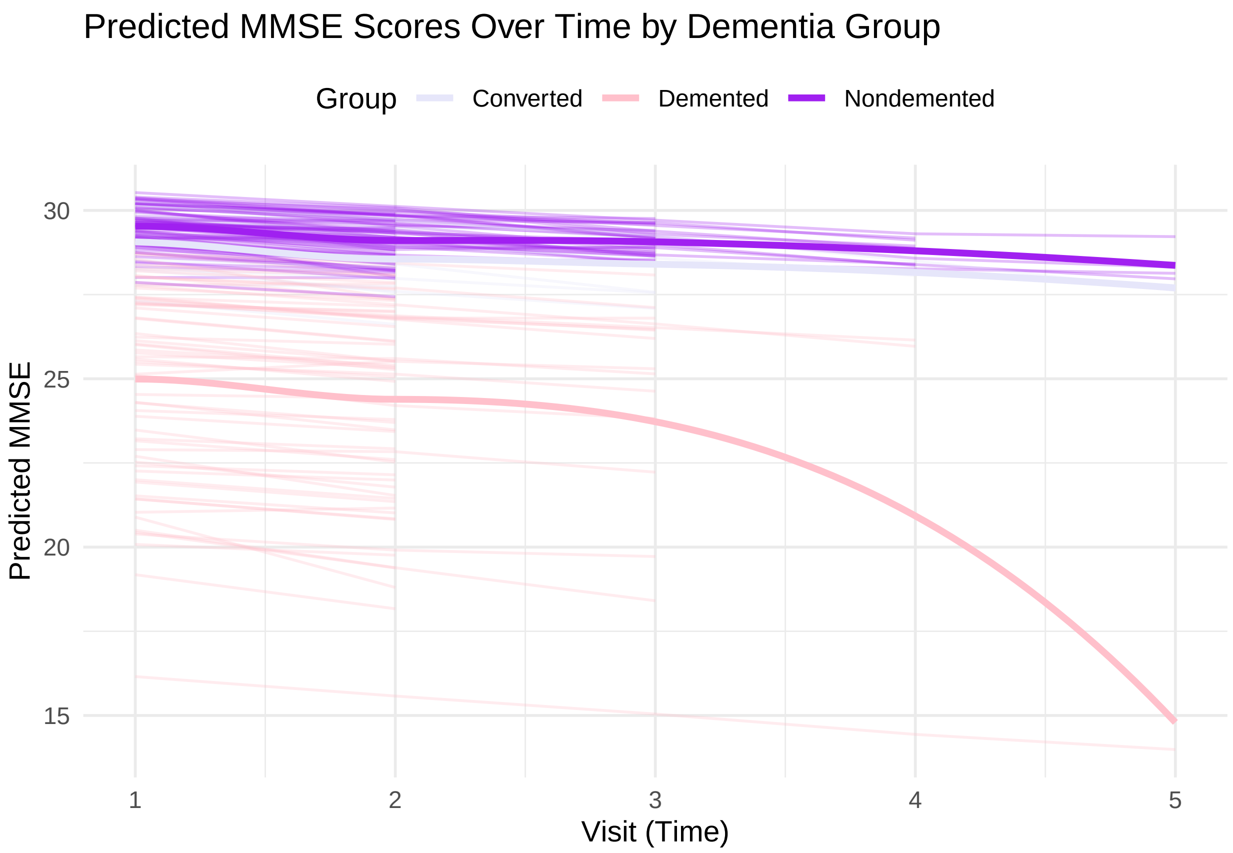
Plotting a Mixed-Effect Model

```
oasis_longitudinal$predicted_MMSE <- predict(mixed_model, newdata = oasis_longitudinal)
head(oasis_longitudinal$predicted_MMSE)
```

```
## [1] 28.90371 28.21585 25.10179 24.20299 23.82066 28.00915
```

```
plot3 <- ggplot(oasis_longitudinal, aes(x = Visit, y = predicted_MMSE, color = Group)) +
  geom_line(aes(group = Subject_ID), alpha = 0.3) +
  geom_smooth(aes(group = Group), method = "loess", se = FALSE, linewidth = 1.2) +
  labs(title = "Predicted MMSE Scores Over Time by Dementia Group",
       x = "Visit (Time)",
      y = "Predicted MMSE") +
  theme_minimal() +
  theme(legend.position = "top") +
  scale_color_manual(values = c("lavender", "pink", "purple"))

plot3
```

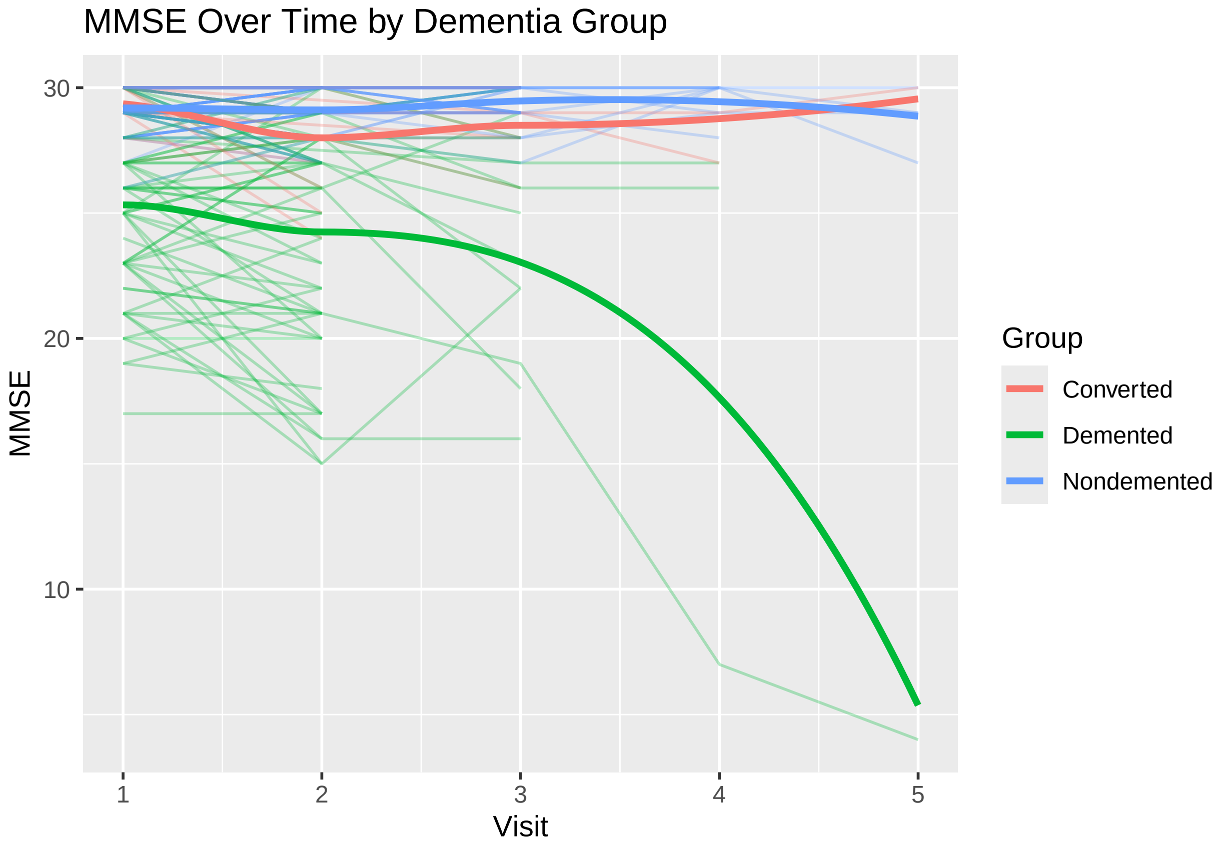
```
## `geom_smooth()` using formula = 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 0.98

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 2.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 8.5134e-17

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 4

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 0.98

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 1.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 0.98

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 1.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 6.8808e-31

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 1
```



Observing Time Trend with Longitudinal Data and Dementia

```
time_trend <- ggplot(oasis_longitudinal, aes(x = Visit, y = MMSE, color = Group)) +
  geom_line(aes(group = Subject_ID), alpha = 0.3) +
  geom_smooth(aes(group = Group), method = "loess", se = FALSE, linewidth = 1.2) +
  labs(title = "MMSE Over Time by Dementia Group", x = "Visit", y = "MMSE")

time_trend
```

```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 0.98

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 2.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 8.5134e-17

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 4

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 0.98
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 1.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 0.98

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 1.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 6.8808e-31

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 1

## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



Results

- The hypothesis was not necessarily confirmed. More analysis is necessary to positively confirm the hypothesis.

- The longitudinal study provides vital information on the progression of Dementia, however, since there is no clear indication of the gap between each visit, I chose to not make a hypothesis for the longitudinal dataset.

Conclusion

- Alzheimer's Disease still requires a significant amount of research to be fully understood. This study employed many computational and statistical tools to further understand trends, however confirmation through scientific study is still necessary.

Data Set Citation

Boysen, J. (2020). *MRI and Alzheimer's Disease.* Kaggle. https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers?resource=download