Detecting Early Alzheimer's Using Longitudinal Data

By

Vindyani Herath

A REPORT

In partial fulfillment of the requirements of

STAT6378 Longitudinal Data Analysis

Department of Mathematics and Statistics

Sam Houston State University

Fall 2020

ABSTRACT

Alzheimer's is a specific brain disease that accounts for 60-80% of dementia cases and it affects memory, thinking and behavior. Sometimes the symptoms grow severe enough to interfere with daily tasks. This study was undertaken with an objective to analyze the factors that can be helpful in detecting early Alzheimer's in patients using longitudinal data. A total of 6 covariates along with an ordinal response variable which gives the clinical dementia rating for 150 randomly selected were used for the analysis. The data were collected for 5 years from subjects aged between 60 and 96. Since the response variable is ordinal, a marginal ordinal regression model and a generalized linear mixed effect model were implemented to assess changes in the odds of a more favorable response over the duration of the study and to determine which covariates have a significant effect on the response variable. Both models showed that mini-mental state examination score has a highly significant effect on clinical dementia rating. Higher values of the score indicates lower clinical dementia rating which means a more favorable response. Further, it can be concluded that at the completion of the study, most of the subjects had the same clinical dementia rating they had at the beginning of the study. Therefore, it can be said that increasing age is not the only factor that influences Alzheimer's disease among patients.

TABLE OF CONTENTS

INTRODUCTION

Alzheimer's is a degenerative brain disease and the most common form of dementia, a general term for memory loss and other cognitive abilities serious enough to interfere with daily life. The greatest known risk factor for Alzheimer's is increasing age and most people with Alzheimer's are 65 and older. However, Alzheimer's cannot be considered as just disease that occurs in old people. Approximately 200,000 Americans under the age of 65 are known to have younger on-set Alzheimer's disease. Even though the memory loss is mild in early stages, individuals with late-age Alzheimer's lose the ability to carry on a conversation and respond to their environment.

It is believed that a combination of lifestyle, environmental and genetic factors influence when Alzheimer's disease begins and how it progresses. Brain imaging via magnetic resonance imaging (MRI) is used for evaluation of patients with suspected Alzheimer's disease. Some studies have suggested that MRI features may predict rate of decline of Alzheimer's disease and may guide therapy in the future.

The objective of this study is to analyze the factors that can be helpful in detecting early Alzheimer's in patients using longitudinal data. The outcome variable of interest is Clinical Dementia rating measured at baseline (year 0), year 1, year 2, year 3 and year 4. This response variable is measured on a 5-level ordinal scale, which means the categories in the variable has a natural order. Therefore, marginal ordinal regression model and generalized linear mixed effect model for an ordinal response is implemented to assess changes in the odds of a more favorable response over the duration of the study and to determine which covariates have a significant effect on the response variable.

Data Description

Data were obtained from the website 'www.kaggle.com' and the MRI related data were generated by the Open Access Series of Imaging Studies (OASIS) project. The dataset consists of longitudinal data of 150 randomly selected subjects aged between 60 and 96.

Description of Variables

Response Variable: Clinical Dementia Rating (CDR)

Clinical Dementia Rating is a 5-point scale used to characterize 6 domains of cognitive and functional performance applicable to Alzheimer's disease and related dementia.

Response Variable	Scale	Description
Clinical Dementia Rating (CDR)	0	**Normal** – No memory loss or slight inconsistent forgetfulness
	1	**Very Mild Dementia** – Benign forgetfulness
	2	**Mild Dementia** – Moderate memory loss
	3	**Moderate Dementia** – Severe memory loss, only highly learned material retained.
	4	**Severe Dementia** – Severe memory loss, only fragments remain

This study contains subjects with Clinical Dementia Rating of 0, 1 and 2.

Covariates	Description
Time	Each subject was scanned 5 times, once every year for 5 years
Age	Age of the subject at the beginning of the study
Gender	M – Male, F – Female
Years of Education	Number of years each subject had education
Socioeconomic Status	Socioeconomic status as assessed by the Hollingshead Index of Social Position and classified into categories from 1(highest status) to 5 (lowest status)
Mini Mental State Examination (MMSE)	30-point questionnaire that is used extensively in clinical and research settings to measure cognitive impairment.
Estimated Total Intracranial Volume (eTIV)	Estimated volume of the cranial cavity as outlined by the supratentorial dura matter or cerebral contour when dura is not

	clearly detectable. An important covariate for volumetric analyses of the brain and brain regions, especially in the neurodegenerative diseases
Normalize Whole Brain Volume	Expressed as a percent of all voxels in the atlas-masked image that are labeled as gray or white matter by the automated tissue segmentation process

METHODOLOGY

Marginal Ordinal Regression Model

Let Y_{ij} denote the ordinal response for the i^{th} subject at the j^{th} occasion with K categories. Marginal model for ordinal response Y_{ij} can be generally defined as a partial proportional odds model. The general marginal ordinal regression model for longitudinal data is given by:

$$log\left[\frac{P(Y_{ij} \leq k)}{1 - P(Y_{ij} \leq k)}\right] = \alpha_k + X'_{ij}\beta$$

Where $k = 1, 2, \dots, K-1$ for the K categories of the ordinal outcome.

This model analyzes the relationship between the response variable and covariates without taking into account the between subject heterogeneity. In this model, the coefficients have a population interpretation rather than an individual one. Therefore, the model is also referred to as population average model. Although the model includes $K-1$ intercepts, it assumes the effects of the predictor variable are the same across the $K-1$ logits. One advantage of this model is, regardless of the number of categories in the response variable, the interpretation if the slope parameter is the same and the slope parameter exponentiated is interpreted as an odds ratio.

Generalized Linear Mixed Effect Model for an Ordinal Response

Generalized linear models can be extended to longitudinal data by allowing a subset of the regression coefficients to vary randomly from one individual to another. These models are known as generalized linear mixed effects models. It is assumed that the random effects have a multivariate normal distribution and responses for any particular individual are independent

observations from a distribution belonging to the exponential family (Multinomial distribution if Y_{ij} is ordinal).

Suppose that Y_{ij} is an ordinal response with K categories $(1, \dots, K)$. The k^{th} cumulative response probability for Y_{ij} depends on fixed and random effects.

$$log\left[\frac{P(Y_{ij} \leq k|b_i)}{P(Y_{ij} > k|b_i)}\right] = \alpha_k + X'_{ij}\beta + Z'_{ij}b_i, \qquad (k = 1, \dots K-1)$$

That is, the conditional cumulative response probabilities are related to the linear predictor by a logit link function.

RESULTS AND DISCUSSION

Preliminary Analysis

Exploratory data analysis is used to determine how the covariates behave with the response variable.

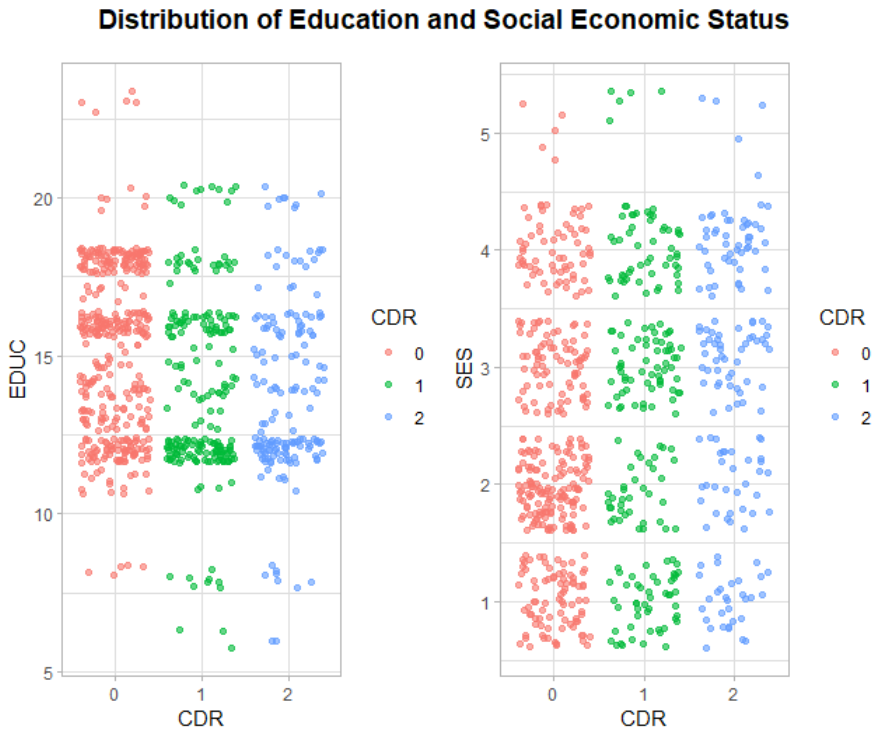

Figure 1 Distribution of Education Social Economic Status with Clinical Dementia Rating

As depicted in Figure 1, there is not any obvious relationship of education and social economic status with clinical dementia rating. Therefore, we can consider removing these two covariates from the model.

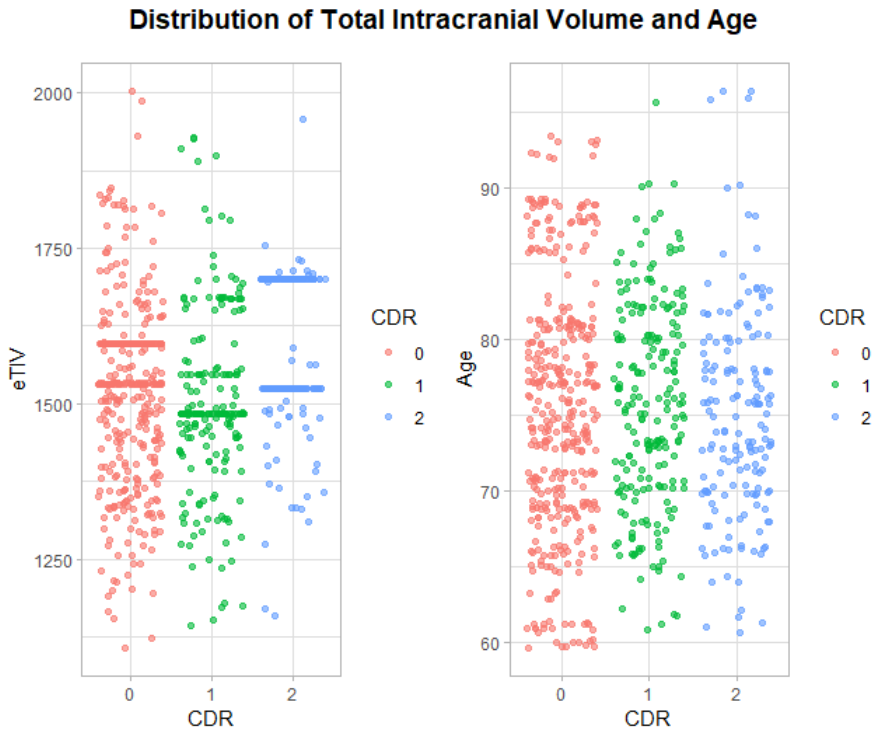

Figure 2 Distribution of Total Intracranial Volume and Age with Clinical Dementia Rating

When considering the distribution of total intracranial volume with clinical dementia rating in Figure 2, it can be observed that intracranial volume is more spread out for patients with no dementia (CDR = 0). For the distribution with age, it can be observed that the oldest subjects in the study have a clinical dementia rating of 1 or 2. In Figure 3, the distribution on Mini-mental state examination with CDR shows that the subjects who were not diagnosed with dementia have a MMSE rating that is concentrated around 27 – 30. MMSE results of subjects diagnosed with dementia seems to be more spread. Subjects who were diagnosed with mild dementia have a very low MMSE rating while some subjects in the same category had very high MMSE rating. For the distribution of normalize whole brain volume with clinical dementia rating, it can be observed that the subjects with a clinical dementia rating of 2 are concentrated below the 0.7 line while other subjects are mode spread out.

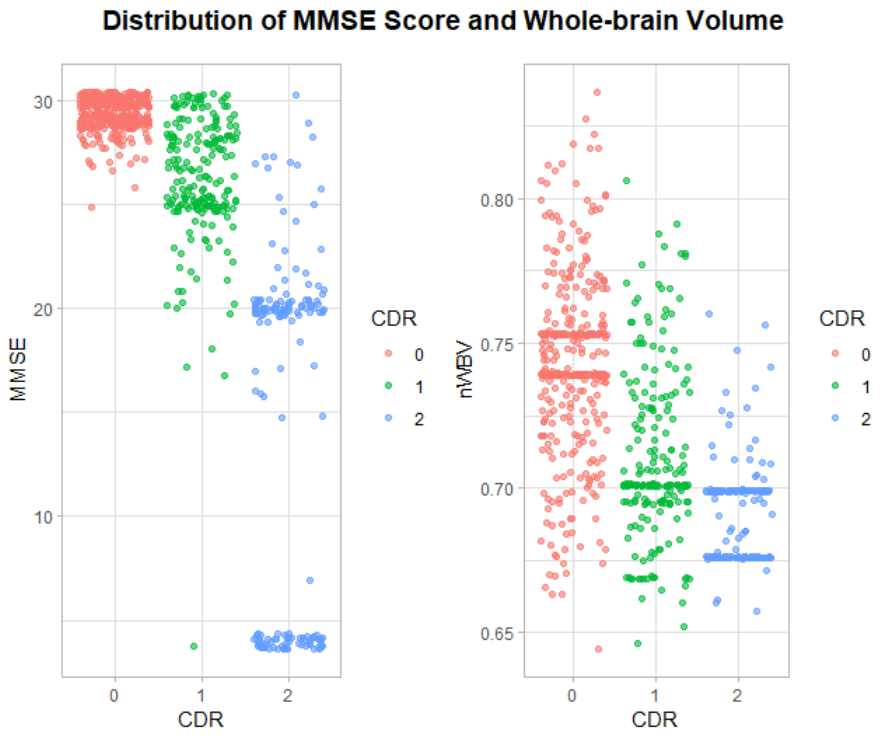

Figure 3 Distribution of Mini Mental State Examination and Normalize Whole Brain Volume with Clinical Dementia Rating

The Ordinal Model without Covariates

The probability of response for each category in the response variable can be found by implementing a model without the covariates.

Model

$$log\left[\frac{P(Y_{ij} \leq k)}{1 - P(Y_{ij} \leq k)}\right] = \alpha_k$$

$$i = 1, 2, \ldots, 150\ subjects$$

$$j = 2, \ldots, 5\ years$$

SAS labels α_k as intercepts where $k = 1, 2$

Based on this model, probability of responses for each category in response variable can be found.

Fitted Model

Table 1 SAS output for the ordinal model without covariates

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**2**	1	-1.2579	0.0880	204.5319	<.0001
Intercept	**1**	1	-0.00533	0.0730	0.0053	0.9418

From table 1,

Intercept 2 = -1.2579 = logit of response in category 2

Intercept 1 = -0.00533 = logit of response in category 1 and 2

Probability of response in category 2 = $\frac{1}{1+\exp(-(-1.2579))} = 0.2213 = 166/750$

Probability of response in category 1 and 2 = $\frac{1}{1+\exp(-(-0.00533))} = 0.4987 = 374/750$

Probability of response in category 1 = $\frac{374}{750} - \frac{166}{750} = \frac{208}{750}$

Probability of response in category 0 = $1 - \frac{374}{750} = \frac{376}{750}$

Marginal Ordinal Regression Model

Model

Let Y_{ij} denote the ordinal response for the ith subject at the jth occasion with K categories

$$log\left\{\frac{P(Y_{ij} \le k)}{P(Y_{ij} > k)}\right\} = \alpha_k + \beta_1(time)_{ij} + \beta_2(Age)_i + \beta_3(Gender)_i + \beta_4(MMSE)_{ij} + \beta_5(eTIV)_{ij} + \beta_6(nWBV)_{ij}$$
$$+ \beta_7(time)_{ij} * (MMSE)_{ij} + \beta_8(time)_{ij} * (nWBV)_{ij}$$

$$i = 1, 2, \dots, 150\ subjects$$
$$j = 0, 1, \dots, 4\ years$$
$$k = 1, 2$$
$$Gender_i = \begin{cases} 1\ if\ Female\ subject \\ 0\ if\ Male\ subject \end{cases}$$

Based on this model, parameters can be estimated using PROC GENMOD function in SAS.

Table 2 SAS output for the marginal ordinal regression model

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		**Estimate**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
Intercept1		-27.2082	5.4584	-37.9064	-16.5099	-4.98	<.0001
Intercept2		-22.4017	5.3760	-32.9385	-11.8649	-4.17	<.0001
time		-14.5014	2.2537	-18.9186	-10.0842	-6.43	<.0001
Age		0.0385	0.0202	-0.0012	0.0781	1.90	0.0574
Gender	F	0.5647	0.2923	-0.0082	1.1375	1.93	0.0534
Gender	M	0.0000	0.0000	0.0000	0.0000	.	.
MMSE		0.7087	0.1068	0.4994	0.9181	6.63	<.0001
eTIV		0.0014	0.0011	-0.0008	0.0035	1.23	0.2176
nWBV		3.1034	5.4013	-7.4829	13.6897	0.57	0.5656
time*MMSE		0.0232	0.0570	-0.0885	0.1349	0.41	0.6840
time*nWBV		19.2936	2.0004	15.3728	23.2143	9.64	<.0001

Two models can be written using the 2 intercepts in table (2).

$$log\left\{\frac{P(Y_{ij} \le k)}{P(Y_{ij} > k)}\right\} = -27.2082 - 14.5014time_{ij} + 0.0385Age_i + 0.5647Gender_i + 07087MMSE_{ij} + 0.0014eTIV_{ij} + 3.1034nWBV_{ij} + 0.0232time_{ij} * MMSE_{ij} + 19.2936time_{ij} * nWBV_{ij} \quad (1)$$

$$log\left\{\frac{P(Y_{ij} \le k)}{P(Y_{ij} > k)}\right\} = -22.4017 - 14.5014time_{ij} + 0.0385Age_i + 0.5647Gender_i + 07087MMSE_{ij} + 0.0014eTIV_{ij} + 3.1034nWBV_{ij} + 0.0232time_{ij} * MMSE_{ij} + 19.2936time_{ij} * nWBV_{ij} \quad (2)$$

Model (1) gives the log of odds of the lowest category compared to other categories while model (2) gives the log of odds for the two lowest categories compared to the remaining category. The model depicts the cumulative log odds of being in the lower numbered categories. Hence, larger values of βx are associated with an increased probability of being in the lower number categories, meaning we get a favorable response (CDR closer to zero) if we have large regression coefficient

values. Further, a positive coefficient means a negative relationship between the covariate and the clinical dementia rating. Therefore, covariates with higher values of coefficients are associated with lower values of the ordinal scale i.e., getting a favorable response (CDR closer to zero).

According to table (2), intercepts, time, Mini-mental state examination and the interaction between time and Normalized whole brain volume are highly significant. It can be said that relative to baseline, the odds of a more favorable response at the end of the study is approximately zero. ($e^{-14.5014*4}$). At the completion of the study, subjects had the same clinical dementia rating they had at the beginning of the study. For each one unit increase in mini-mental state examination (MMSE), cumulative log odds of clinical dementia rating change by 0.7087, while holding other variables constant. Therefore, it can be said that higher values of MMSE are associated with lower clinical dementia rating. The odds of clinical dementia rating is greater among female subjects ($e^{0.5647} = 1.76$) compared to male subjects. Female subjects are approximately twice as likely to have a low clinical dementia rating compared to male subjects.

Generalized Linear Mixed Model for an Ordinal Response

Random Intercept-Slope Model

$$log\left\{\frac{P(Y_{ij} \leq k)}{P(Y_{ij} > k)}\right\} = \alpha_k + \beta_1(time)_{ij} + \beta_2(Age)_i + \beta_3(Gender)_i + \beta_4(MMSE)_{ij} + \beta_5(eTIV)_{ij} + \beta_6(nWBV)_{ij}$$
$$+ \beta_7(time)_{ij} * (MMSE)_{ij} + \beta_8(time)_{ij} * (nWBV)_{ij} + b_{1i} + b_{21}(time)_{ij}$$

$$i = 1, 2, \dots, 150\ subjects$$
$$j = 0, 1, \dots, 4\ years$$
$$k = 1, 2$$
$$Gender_i = \begin{cases} 1\ if\ Female\ subject \\ 0\ if\ Male\ subject \end{cases}$$

Based on this model, parameters can be estimated using PROC GLIMMIX function in SAS.

Fitting both random intercept model and random intercept-slope model and comparing the Akaike Information Criterion values shows that the best fitted model for the data set is the random intercept-slope model.

Table 3 SAS output for the random intercept and slope model

Solutions for Fixed Effects							
Effect	**CDR**	**Gender**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	0		-26.6089	6.1611	149	-4.32	<.0001
Intercept	1		-20.8880	6.0566	149	-3.45	0.0007
time			-12.3192	2.3932	149	-5.15	<.0001
Age			0.01183	0.02520	442	0.47	0.6388
Gender		F	0.6314	0.3752	442	1.68	0.0931
Gender		M	0	.	.	.	.
MMSE			0.8404	0.08986	442	9.35	<.0001
eTIV			0.000137	0.001120	442	0.12	0.9026
nWBV			2.6129	6.6338	442	0.39	0.6939
time*MMSE			-0.02259	0.02998	442	-0.75	0.4514
time*nWBV			17.9713	3.4153	442	5.26	<.0001

The estimated parameters are almost same as the marginal model. Intercept terms correspond to the 2 cumulative logits that are defined on the clinical dementia rating in the order shown. Intercept 1 is the log of odds of clinical dementia rating 0 vs. 1 while intercept 2 is of rating 0,1 and 2. Conditioned on random effects, the odds of clinical dementia rating decreasing is approximately zero ($e^{-12.3192} \approx 0$) for each year. Given 2 subjects have the same random effect, with one unit increase in mini mental state examination, the log of odds of clinical dementia rating increases by 0.8404. Further, given two subjects have the same random effect, the odds of clinical dementia rating is greater among female group ($e^{0.6314} = 1.88$) as compared to the male group.

CONCLUSION

Alzheimer's is a brain disease which causes memory loss and interferes with daily life of patients. It is important to identify the disease at an early age. In this study, factors that can be helpful in identifying Alzheimer's were analyzed using longitudinal data. Marginal model for an ordinal response variable and generalized linear mixed model for an ordinal response were analyzed. From

both models, it can be concluded that mini-mental state examination score has a highly significant effect on clinical dementia rating. Higher values of the score indicates lower clinical dementia rating and that means a favorable response closer to no dementia. Both models suggested that the odds of clinical dementia rating is greater among female subjects than the male subjects, which means that female subjects are more likely to have a lower clinical dementia rating. Finally, it can be concluded that at the completion of the study, most of the subjects had the same clinical dementia rating they had at the beginning of the study. Therefore, it can be said that increasing age is not the only factor that influences Alzheimer's disease among patients.

APPENDIX

ANALYSIS USING SAS

Importing Data

```
/*Importing data */
data MRI;
infile "/folders/myfolders/sasuser.v94/Longitudinal/MRI.dat" DELIMITER='09'x;
input ID Group $ time Gender $ Age EDUC SES MMSE CDR eTIV nWBV ASF;
run;
data MRI;
        set MRI;
        IF time=1 THEN time = 0;
        IF time=2 THEN time = 1;
        IF time=3 THEN time = 2;
        IF time=4 THEN time = 3;
        IF time=5 THEN time = 4;
proc print;
run;
proc means data= MRI maxdec = 2 n mean var nway;
        var MMSE ASF eTIV nWBV Age ;
        class CDR time;
        output out=outmean mean=mean;
 run;
```

Model without covariates

```
/* Model without Covariates */
PROC LOGISTIC data=MRI DESCENDING;
model CDR = ;
```

Multinomial Logistic Regression Model

```
/*Multinomial Logistic Regression Model */

PROC LOGISTIC DATA=MRI;

CLASS ID Gender (REF = "M") / PARAM= REF;

MODEL CDR=time Age Gender MMSE eTIV nWBV time*MMSE time*nWBV / LINK=CLOGIT
SCALE=NONE AGGREGATE RSQ LACKFIT;

RUN;
```

Marginal ordinal Regression Model

```
/* Marginal Model */

PROC genmod data=MRI;

        CLASS ID Gender;

        MODEL CDR=time Age Gender MMSE eTIV nWBV time*MMSE time*nWBV /
dist=multinomial link=cumlogit type3 wald;

        repeated subject = id / type = ind;

run;
```

Random Intercept Model

```
/* Random-intercept Model */

PROC GLIMMIX DATA=MRI METHOD=QUAD NOCLPRINT;

        CLASS ID Gender;

        MODEL CDR=time Age Gender MMSE eTIV  nWBV time*MMSE time*nWBV
/SOLUTION DIST=MULTINOMIAL LINK=CUMLOGIT;

        RANDOM INTERCEPT /SUBJECT=ID;

RUN;
```

Random Intercept and Slope Model

```
/* Random intercept & slope model */

PROC GLIMMIX DATA=MRI METHOD=QUAD NOCLPRINT;

        CLASS ID Gender;

        MODEL CDR=time Age Gender MMSE eTIV  nWBV time*MMSE time*nWBV
/SOLUTION DIST=MULTINOMIAL LINK=CUMLOGIT;

        RANDOM INTERCEPT time /SUBJECT=ID;

RUN;
```

R CODES FOR GRAPHS

```
MRI <- read.csv("C:/Users/vindy/Google Drive/SHSU/Longitudinal Data Analysis/Project/MRI_data.csv")

attach(MRI)

library(ggplot2)

library(tidyverse)

x <- MRI %>%
  select(EDUC, CDR, Gender) %>%
  mutate(CDR = as.factor(CDR)) %>%
  ggplot(aes(x = CDR, y = EDUC)) +
  geom_jitter(aes(col = CDR), alpha = 0.6) +
  theme_light() ;x

y <- MRI %>%
  select(SES, CDR, Gender) %>%
  mutate(CDR = as.factor(CDR)) %>%
  ggplot(aes(x = CDR, y = SES)) +
  geom_jitter(aes(col = CDR), alpha = 0.6) +
  theme_light() ;y

library(cowplot)

p <- plot_grid(x, y)
```

```
title <- ggdraw() + draw_label("Distribution of Education and Social Economic Status", fontface='bold')
plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))


x <- MRI %>%
  select(MMSE, CDR, Gender) %>%
  mutate(CDR = as.factor(CDR)) %>%
  ggplot(aes(x = CDR, y = MMSE)) +
  geom_jitter(aes(col = CDR), alpha = 0.6) +
  theme_light() ; x

y <- MRI %>%
  select(nWBV, CDR, Gender) %>%
  mutate(CDR = as.factor(CDR)) %>%
  ggplot(aes(x = CDR, y = nWBV)) +
  geom_jitter(aes(col = CDR), alpha = 0.6) +
  theme_light() ; y

p <- plot_grid(x, y)
title <- ggdraw() + draw_label("Distribution of MMSE Score and Whole-brain Volume", fontface='bold')
plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))


x <- MRI %>%
  select(eTIV, CDR, Gender) %>%
  mutate(CDR = as.factor(CDR)) %>%
  ggplot(aes(x = CDR, y = eTIV)) +
  geom_jitter(aes(col = CDR), alpha = 0.6) +
  theme_light() ;x
```

```
y <- MRI %>%
  select(ASF, CDR, Gender) %>%
  mutate(CDR = as.factor(CDR)) %>%
  ggplot(aes(x = CDR, y = Age)) +
  geom_jitter(aes(col = CDR), alpha = 0.6) +
  theme_light() ;y


p <- plot_grid(x, y)
title <- ggdraw() + draw_label("Distribution of Total Intracranial Volume and Age", fontface='bold')
plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))
```

REFERENCES

1. Applied longitudinal analysis / Garrett M. Fitzmaurice, Nan M. Laird, James H. Ware
2. Silva, Nívea & Amorim, Leila & Colosimo, Enrico & Heller, Léo. (2016). Modeling ordinal longitudinal outcomes: an applied perspective of marginal and conditional approaches. Chilean journal of statistics. 7. 51-68.
3. Analyzing longitudinal data and use of the generalized linear model in health and social sciences. / Arnau, J., Bono, R., Bendayan, R., & Blanca, M. J. (2015). Quality & Quantity, 50(2), 693–707.
4. Hedeker, D. (2008). Multilevel models for ordinal and nominal variables. In J.de Leeuw & E.Meijer (Eds.), Handbook of Multilevel Analysis. Springer, NewYork.
5. Hedeker,D.(2005).Generalizedlinearmixedmodels.InB.Everitt&D.Howell(Eds.),EncyclopediaofStatisticsinBehavioralScience.Wiley.
6. Kiernan, K. (2018). Insights into Using the GLIMMIX Procedure to Model Categorical Outcomes with Random Effects.
7. Nick R. Parsons, Matthew L. Costa, Juul Achten, Nigel Stallard, Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R, Computational Statistics & Data Analysis, Volume 53, Issue 3, 2009,
8.