Alzheimer Prediction - Multilevel Regression Analysis

Hao He

2022-11-30

Abstract

Early detection of Alzheimer's Disease is essential for dementia prevention. The use of MRI imaging seems promising for early detection. However, with MRI scans, it is still hard for doctors to assess whether a patient's condition surely can be diagnosed with dementia at as early as possible. This analysis aims to figure out if it is possible to predict who has Alzheimer's by using the multilevel logistic model after some signs are noticed and which factor(s) would be influential. The result suggests that MMSE test result, age in years and years of education are important factors.

Introduction

Alzheimer's disease (AD), as the most common trigger of dementia among older people, will shrink the size of a brain and cause slow decline in memory, thought processes, language and functioning. After AD is diagnosed, the symptoms gets worse, and no existing drugs in market can cure the disease. Thus, I'm curious about what patient factors are most related to whether a patient has Alzheimer's and therefore more preventive procedures could be planned for the future to slow down the progression of dementia. As a result, patients can hopefully extend their life expectancy.

The clinical data I found is longitudinal data, so it's natural to have a group level `patient` given each patient was scanned and measured on each visits. The individual-level observations including patient's social demographics and other clinical test results are nested within the patient - level. Since the multilevel model is highly effective to understand the impacts of mixed effects and make prediction, I built a multilevel logistic model to reflect the nested data structure and to predict if a patient has AD and how that varies between different patients. In this analysis, how I choose variables and modeling will be elaborated in Methods section. The Result section summarizes the interpretation of final model's results and the limitations and possible improvements is reported in Discussion section.

Methods

The Longitudinal MRI Data in Nondemented and Demented Older Adults dataset in this analysis can be found on Kaggle, and it was originally published on the Open Access Series of Imaging Studies (OASIS) website as OASIS-2 [1]. In this dataset, 150 subjects aged from 60 to 96 are repeatedly measured over a period of time and information collected include group, number of visit, MR delay, gender, handedness, age, education, social economic status (SES), MMSE, CDR, eTIV, nWBV and ASF. Below is a data description for each variable [2].

Data Description

Subject ID	Identification of patients
MRI ID	Identification of MRI scans
Group	Demented or Nondemented
Visit	The visit number
MR Delay	The number of days between two medical visits
M/F	Gender
Hand	Dominant Hand
Age	Age in years
EDUC	Years of Education
SES	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) 1 = upper, 2 = upper middle, 3 = middle, 4 = lower middle, 5 = lower)
MMSE	Mini Mental State Examination (range from 0 (worst) to 30 (best))
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor

EDA

After initial screening of data, I found there are 21 missing values in this dataset, I kept them for EDA part to avoid mishandling the data and miss any patterns. Based on the OASIS FACT SHEET [2], I knew that all participants with dementia (CDR >0) were diagnosed with probable AD. Therefore, I chose `CDR` as the response variable for prediction of AD and converted it to a binary variable `Response` (0 = Nondemented, 1 = Demented).

I plotted each variable's distribution to make sense of what variables are more reliable to put into my model in terms of normality. In **Figure 1** (see Appendix), the majority of the variables have a distribution not too far way from from normal distribution, however the MMSE and EDUC variable may need a logarithmic transformation. The MMSE variable looks left-skewed and has a heavy tail while EDUC has a lot of peaks.

According to the dementia status, patients are categorized into three groups, demented, nondemented, and converted. By examining the converted group, I found 14 patients are converted from nondemented to demented after 1 st visit, approximately half of them are detected with very mild dementia (see **Table 1** in Appendix). Since patient's visit is separated by at least one year[1], to reduce speculation around when these

converted patients started to have AD, I grouped the converted and demented patients together as the demented group while nondemented group remains the same. This is helpful to understand any potential patterns because the outcome variable is binary.

Social-Demographics Variable

The relationship between social- demographics variable and dementia are explored in **Figure 2** and patterns are summarized below.

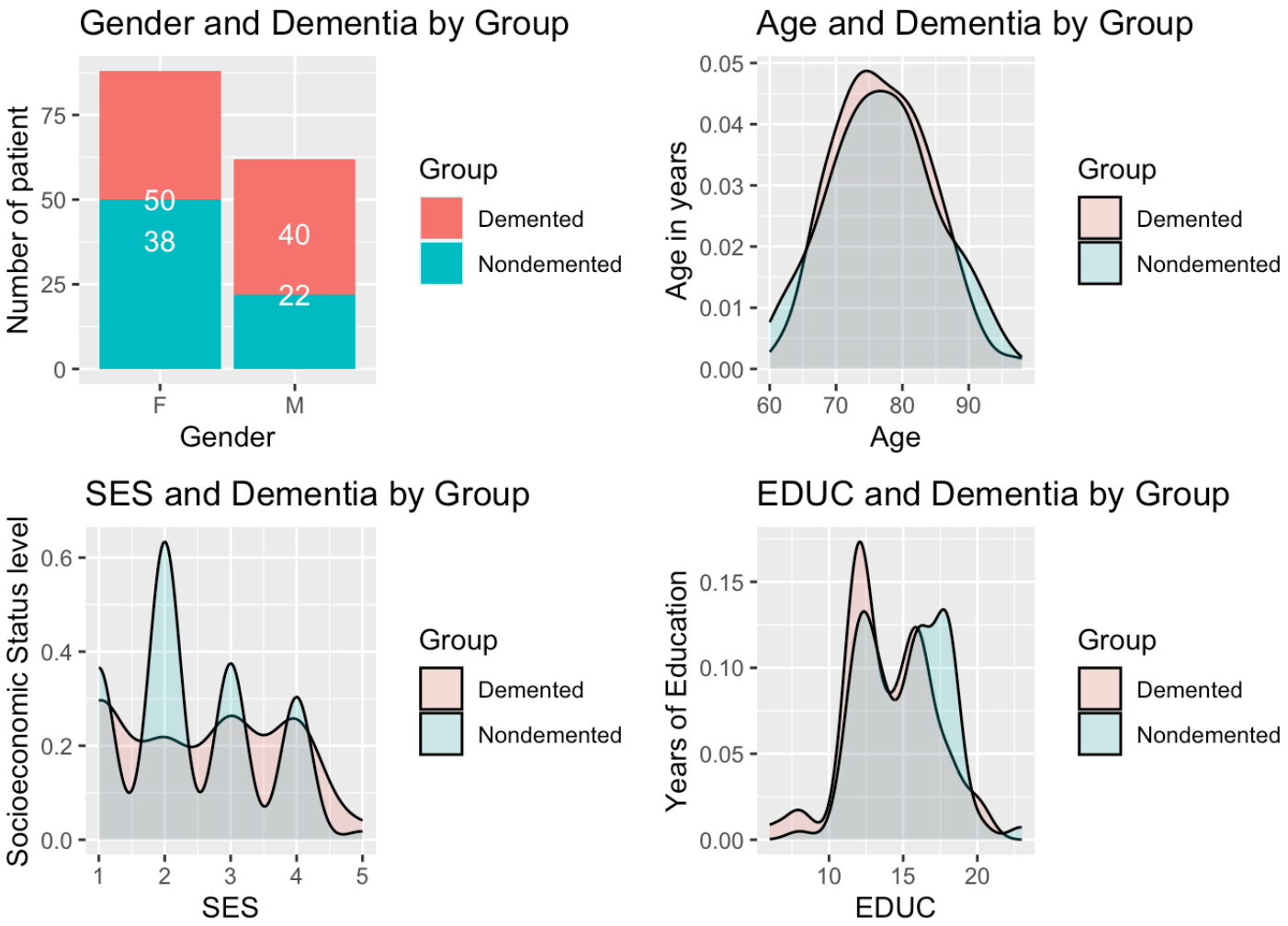

Figure 2. Social-Demograhpics and Dementia

- `Gender` variable The bar chart shows there are less women patients in the demented group, given we have 88 females and 62 males in this dataset.
- `Age` variable I assessed the relationship between age and dementia with my guess being that this kind of disease starts in people's 80s and 90's. It shows that demented group has a higher proportion of patients aged from 65-85 than nondemented group. I also notice that in this dataset the nondemented group has more 90-year-old patients and I assume this is attributed to the fact that dementia reduces life expectancy.
- `SES` variable Demented patients tend to be in a relatively lower social economic status than nondemented patients, especially in the upper middle socioeconomic status.

- `EDUC` variable Demented group has more patients that are less than high school graduates while nondemented group has more patients that are college graduates and even beyond college level.

Clinical Variable

In Figure 3, I explored all clinical variables representing brain diagnostic test to see how each of them is related to dementia. The takeaways from **Figure 3** are

- Demented patients tend to have a higher concentration in ASF measurement between 1.1 and 1.25 and a higher concentration in eTIV values between 1375 and

1500. Besides, `ASF` and `eTIV` seems to be negatively related to some extent.

- Demented patients tend to have a much lower MMSE score than nondemented patients. This may indicate that the lower the MMSE score, the higher the likelihood of patients diagnosed with Alzheimer's.

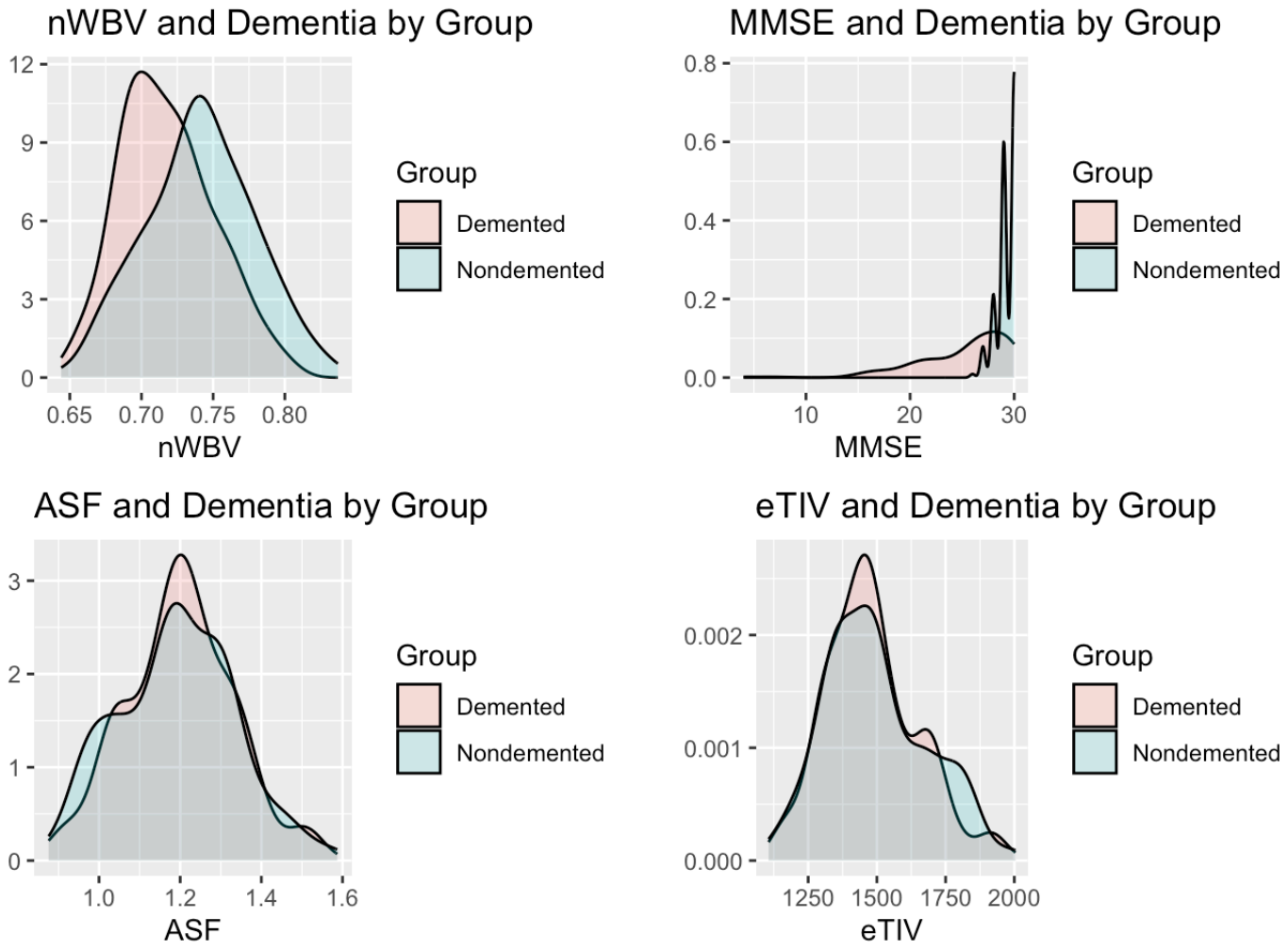

Figure 3. Different brain diagnostic test measurements and Dementia

Below are some additional insights I gained about the relationship between other predictors in this dataset indicated in **Figure 4**.

- `SES` and `EDUC` These two variables has a strong negative correlation (see **Figure 5** in Appendix) and this feature is shared in the correlation between `eTIV` and `ASF` variable, given 0.7 as my baseline of high correlation. Considering this, I decided not to keep the `SES` variable in my model as it's more

categorical and hard to interpret.

- `eTIV` and `ASF` There is a clear linear trend in the **Figure 4**. The reason why `eTIV` and `ASF` are highly correlated is that atlas normalization equates head size causing the ASF to be proportional to TIV [3]. I decided not to keep the `ASF` variable in my model.
- `nWBV` and `Age` In general, the older the people is, the smaller the brain volume is. Demented patients have a smaller size of brain compared to nondemented patients as expected.

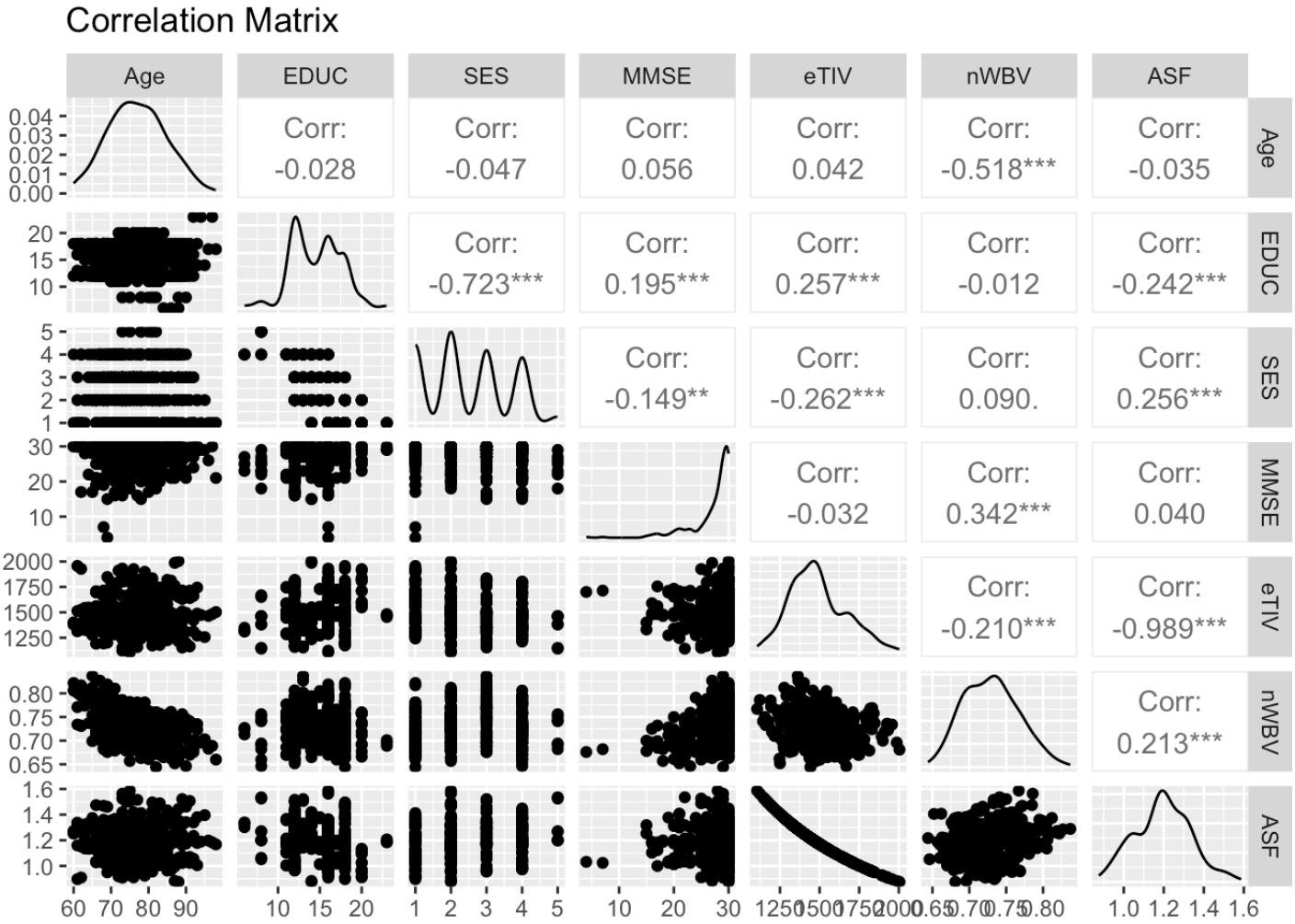

Figure 4. Correlation Matrix

Data Preprocessing

Each patient has at least two or more visits, so I kept data with total visits less than 3 to make sure I don't misinterpret the diagnosis and skew the data because my goal is trying to predict AD at a earlier time. As this dataset only includes 150 subject and the NAs are are Missing at Complete Random, I decided to impute the NAs by mean (MMSE variable) and median (SES variable) instead of remove them. Among all 13 variables, I removed three unnecessary columns, they are `MR delay` , `MRI ID` , and `Hand` because they are not informative for estimating dependent variable `Response` . After these steps, I got a 294-observations dataset with 1 dependent variable `Response` and 11 predictors.

Variable Selection

Based on EDA results, I have 7 predictors for variable selection, they are `Gender` , `MMSE` , `EDUC` , `nWBV` , `eTIV` , and `Age` . Among them, I decide to only keep `MMSE` , `nWBV` , `eTIV` , `EDUC` , and `Age` as predictors. In addition, as I mentioned before in EDA section, MMSE has a skewed heavy-tailed distribution and a logarithmic transformation could help to solve this issue, so I transformed the `MMSE` to a log scale.

Model Fitting

The effect of aforementioned predictors would be natural to differ by each patient, so I first constructed a varying-intercept multilevel as below.

```
mod1 <- stan_glmer(Response ~ (1| pid) + Age + EDUC + log(MMSE) + nWBV + eTIV,
                   data = p1, family = binomial(link = "logit" ))
```

However, the estimated coefficient of `log(MMSE)` is -38.16 with a standards error of 6.45, which suggests the log transformation is not a good decision. So I decided to continue with `MMSE` at the original scale. I also noticed that `eTIV` has near zero coefficients, so I dropped the `eTIV` variable and continue with the model below.

```
mod2 <- stan_glmer(Response ~  Age + EDUC + MMSE + nWBV+ (1| pid), data = p1,
                   family = binomial("logit"))
```

With the standard error of `nWBV` being 16.5 based on the model summary in Appendix, it is impossible to claim that `nWBV` is an important factor for prediction. And the reason of this big standard error is probably because `nWBV` is correlated with Age and MMSE.

But the rest of the variables still seems important in the same way. So the final model and its summary is as below.

```
mod3 <- stan_glmer(Response ~ (1| pid) + Age + EDUC + MMSE, data = p1,
                   family = binomial(link = "logit" ))
```

```
## stan_glmer
##  family:       binomial [logit]
##  formula:      Response ~ (1 | pid) + Age + EDUC + MMSE
##  observations: 294
## ------
##             Median  MAD_SD
## (Intercept) 46.6019  8.7424
## Age         -0.0522  0.0529
## EDUC        -0.2753  0.1515
## MMSE        -1.3967  0.2366
##
## Error terms:
##  Groups Name        Std.Dev.
##  pid    (Intercept) 3.4
## Num. levels: pid 150
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Below is the first 6 random effects parameters of patients.

```
##           (Intercept)
## OAS2_0001       -2.30
## OAS2_0002        2.02
## OAS2_0004       -1.94
## OAS2_0005        0.59
## OAS2_0007        2.41
## OAS2_0008       -1.16
```

Result

Interpretation

The interpretation of the this model needs to fix the varying effect - patient, which means the coefficients can be only interpreted as odds ratio when the patient is the same.

The interpretation of fixed effects here should be similar to logistic model. For example, for `Age` , a one unit increase in years of age is associated with a 0.05 unit decrease in the expected log odds of having Alzheimer's. It may seems not straightforward at the first glance, so I transformed the estimate to a odds ratio. The odds ratio here is the conditional odds ratio for a patient holding `EDUC` and `MMSE` constant as well as for a patient with the same doctor, or doctors with identical random effects. The exponent of 0.05 is greater than 1, it indicates that as the age of a patient increases the predicted probability of a patient is diagnosed is lower than a patient is nondemented. This contradicts with our findings in EDA, this may suggest the collinearity exist.

For `EDUC`, a one unit increase in years of education is associated with a 0.28 unit decrease in the expected log odds of having Alzheimer's. It can also be interpreted qualitatively that people who are less educated are expected to have higher log odds of being diagnosed with Alzheimer's than people who are well-educated. For `MMSE`, a one unit increase in `MMSE` score is associated with a 1.4 unit decrease in the expected log odds of having Alzheimer.

Model Checking

Below is the posterior density check plots, it indicates the model fit is okay with acceptable variations compared to the observed data we have.

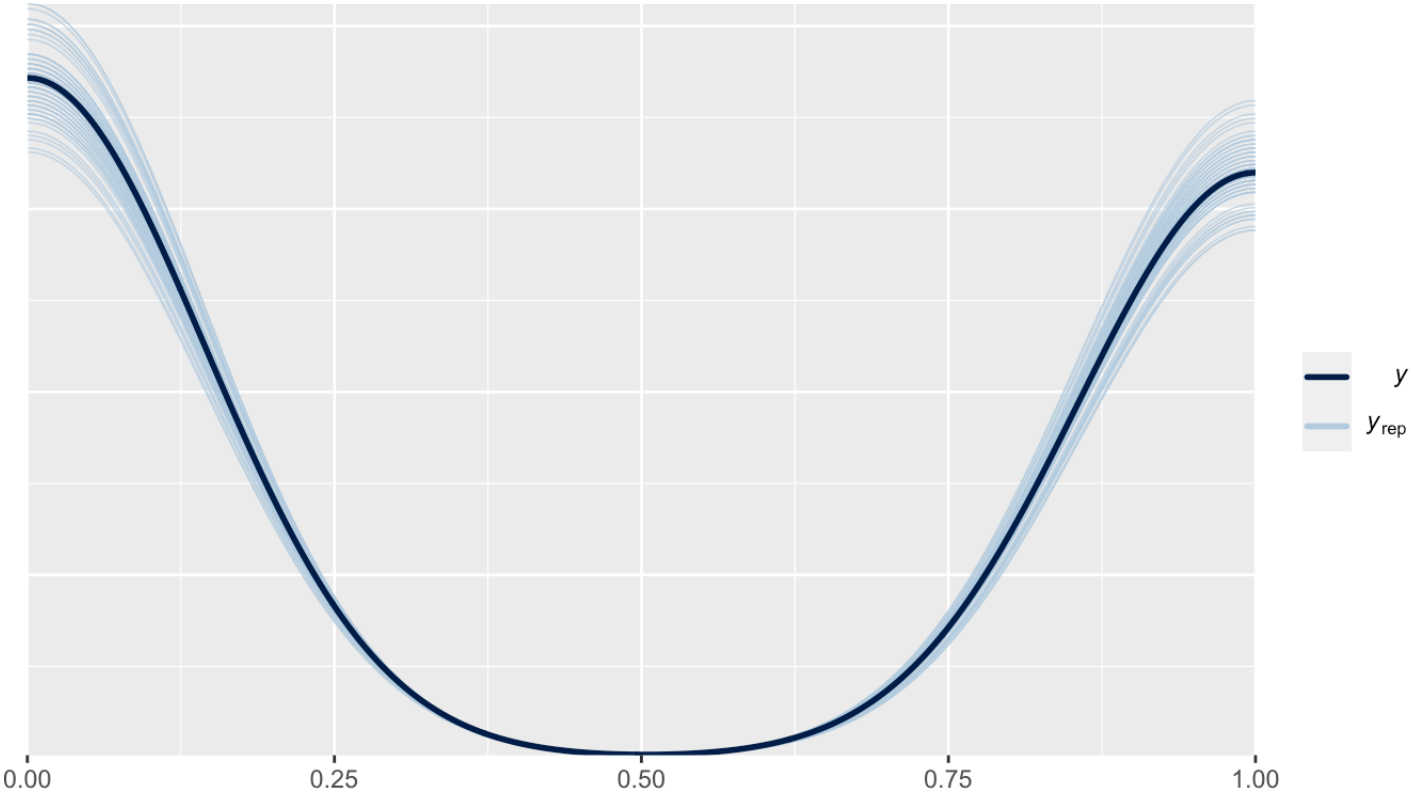

Figure 6. Posterior Predictive Check

However, the are some issues with model checking. Figure 7 below indicates the checking on normality, collinearity, residuals and outliers of this model. The normality seems acceptable although there are some points on tail and head falls off the line and this may be attributed to the outliers shown in the influence plot which is above the Q-Q plot. Also, based on the residuals, we know that the predictive performance at the most of time is okay but with relatively lower predictive accuracy. Furthermore, it is normal to see collinearity happens because this is a longitudinal data and the Alzheimer's is closely related with age.

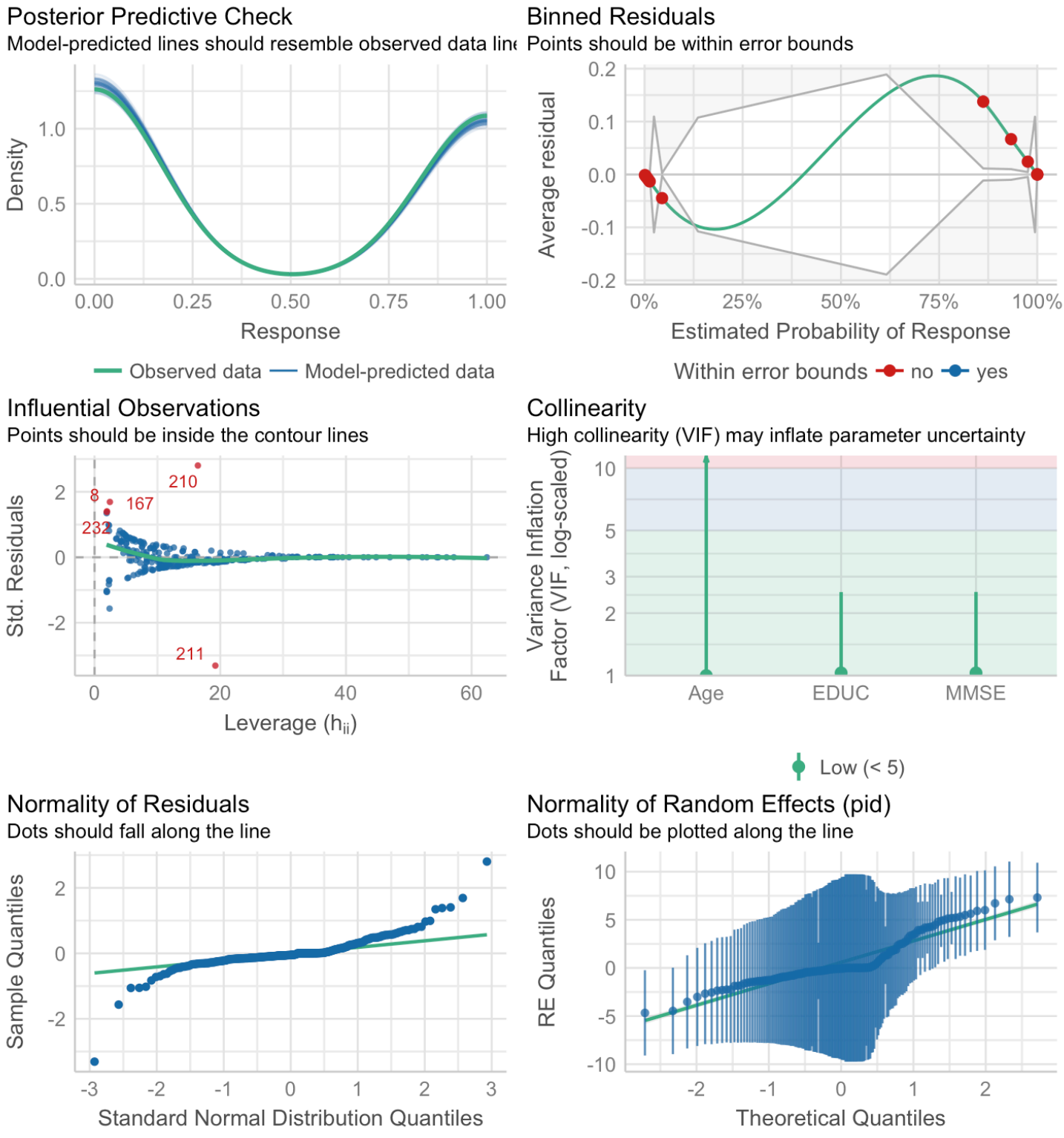

Figure 7. model check

Discussion

In this analysis, I built a multilevel logistic model to predict if a patient has Alzheimer's using three influential predictors, `Age`, `EDUC`, `MMSE`. I didn't consider any interaction term here given it's rank deficient if I only stick to the three individual-level predictors in the final model. This would make the random effects

unidentifiable. I was able to figure out the overall negative relationship between my predictors and the dependent variable.

This analysis do have several limitations and the possible improvements regards to the issues are demonstrated below.

- Because this is an observational study and there is no treatment involved, it would be hard to make causal inference, especially in a multilevel logistic regression.
- The model is built under the assumption that patient visit data are kept at only 2 visits. It may be better to do a performance comparison using data that contains all the visits.
- The method to deal with correlated variables should be considered more carefully. A PCA analysis may be useful to improve the performance of our model because we have several correlated variables.

Reference

[1] “How Is Alzheimer’s Disease Diagnosed?” National Institute on Aging, U.S. Department of Health and Human Services, https://www.nia.nih.gov/health/how-alzheimers-disease-diagnosed (https://www.nia.nih.gov/health/how-alzheimers-disease-diagnosed).

[2] Marcus, Daniel S, et al. “Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults.” Journal of Cognitive Neuroscience, U.S. National Library of Medicine, Dec. 2010, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2895005/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2895005/).

[3] “Oasis Brains.” OASIS Brains - Open Access Series of Imaging Studies, http://www.oasis-brains.org/#data (http://www.oasis-brains.org/#data).

[4] “Home.” OARC Stats, https://stats.oarc.ucla.edu/r/dae/mixed-effects-logistic-regression/ (https://stats.oarc.ucla.edu/r/dae/mixed-effects-logistic-regression/).

Appendix

Dataset Additional Information

This set consists of a longitudinal collection of 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 72 of the subjects were characterized as nondemented throughout the study. 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease. Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

Univariate Analysis

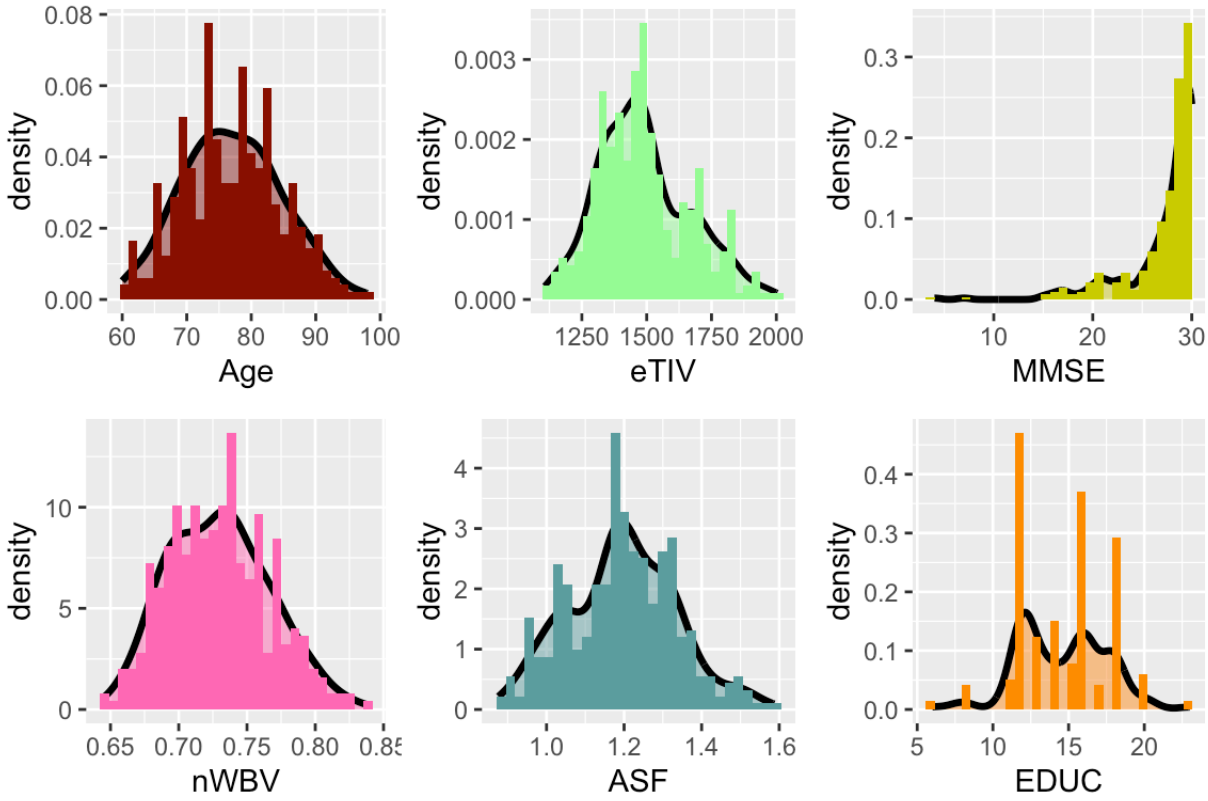

Figure 1. Variable Distributions Plot

EDUC and SES

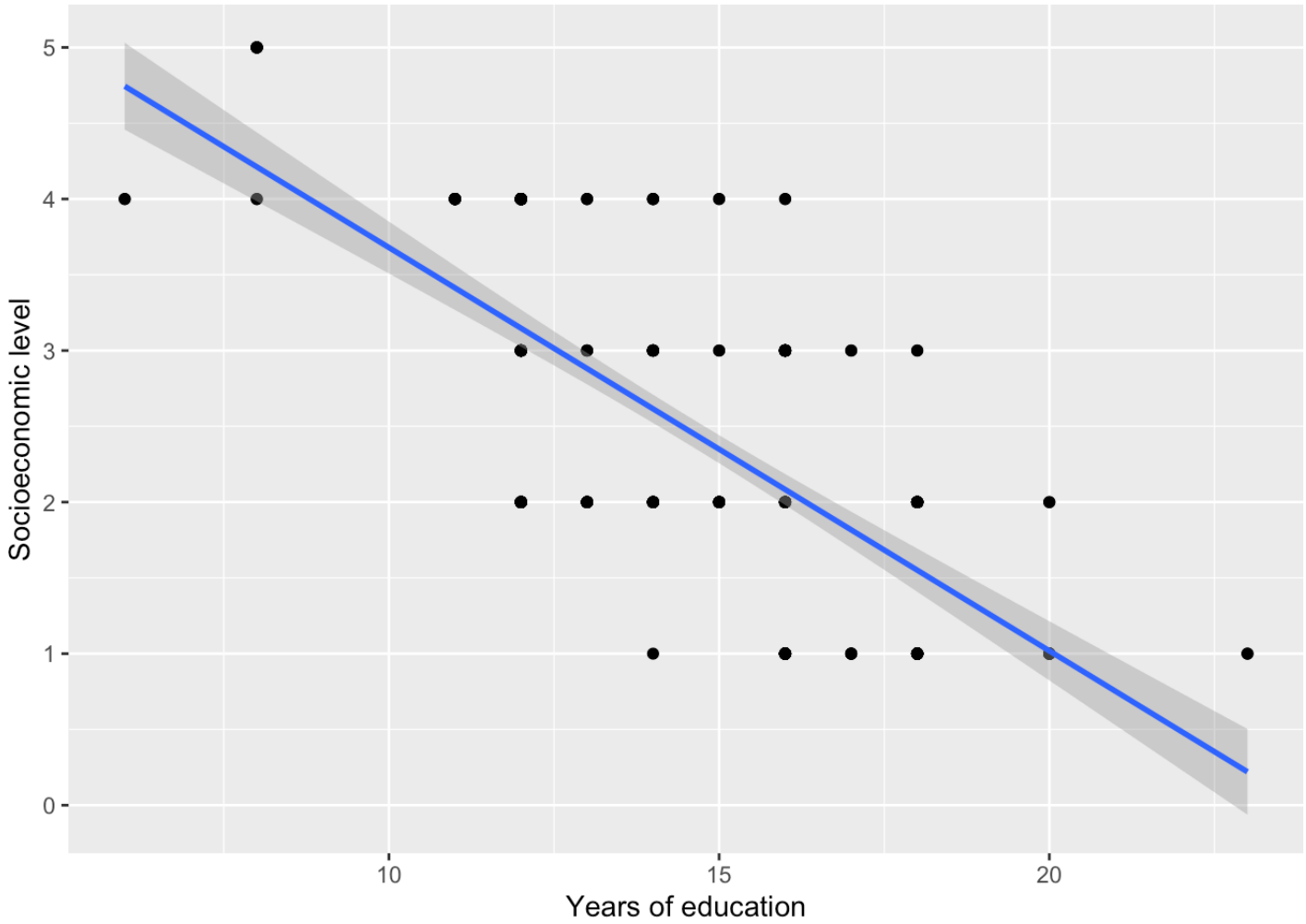

Figure 5. Relationship between EDUC and SES

Converted Group Table

Table 1. Vist and CDR Summary

Visit	CDR	n
1	0.0	13
1	0.5	1
2	0.0	4
2	0.5	8
3	0.0	1
3	0.5	7
4	0.5	2
5	0.5	1

Model Summary

```
## stan_glmer
##  family:       binomial [logit]
##  formula:      Response ~ (1 | pid) + Age + EDUC + MMSE + nWBV
##  observations: 294
## ------
##             Median MAD_SD
## (Intercept)  97.1   19.6
## Age          -0.2    0.1
## EDUC         -0.3    0.2
## MMSE         -1.4    0.2
## nWBV        -51.9   15.7
##
## Error terms:
##  Groups Name        Std.Dev.
##  pid    (Intercept) 3.3
## Num. levels: pid 150
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```