Dementia Patient Demographics Analysis

Greg Pollock

12/14/2020

Executive Summary

Dementia is a growing problem among older adults. This study uses the Clinical Dementia Rating (CDR) as a metric for measuring a person's impaired mental capabilities. Moreover, the analysis in this study shows patients over 60 had an affected CDR at the alpha=0.05 level when age, sex and college attendance were considered. On average, males were 1.9 times more likely to have dementia while college attendees were about half as likely to have or develop it. In addition, as an individual's age is increased year after year, they were, on average, 2.5 times more likely to develop dementia.

Table of Contents

Executive Summary 1
1. Introduction 3
2. Exploratory Data Analysis 3
    2.1 Data Visualization 3
3. Methodology 5
4. Results 6
    4.1 Final Model 7
5. Conclusion 8
References 9
R Code 10
Dataset Exerpt 11

1. Introduction

Dementia is the general term for a person's chronic impaired ability to remember, think or make decisions when doing everyday tasks and is not a normal part of aging. There were an estimated 5 million U.S. adults with dementia in 2014. This number is projected to grow to 14 million by 2060 according to the Centers for Disease Control and Prevention. Clinical Dementia Rating (CDR) is a test performed by medical professionals consisting of simple interview questions meant to target a patient's mental faculties. Following an interview, a score of either 0.0, 0.5, 1.0, 1.5, or 2.0 is assigned to the patient indicating the severity of impairment. It has been shown to be effective in measuring an individual's level of dementia (Rockwood, K. et al.).

My interest is in which patient demographics correlate with having dementia. In this analysis I utilized a generalized linear model to determine the relationship age, gender and college attendance had on having a non-zero CDR (having at least mild dementia). Specifically, a logistic regression was done to measure these effects on the odds of having a non-zero CDR. The details of this method are included later on in the report. In this analysis, an alpha level of $\alpha = 0.05$ was used.

2. Exploratory Data Analysis

Data from the OASIS Project was used. This project's purpose is to provide MRI imaging data from both demented and nondemented patients to the scientific community for study and prediction of Alzheimer's Disease (OASIS Brains).
The specific dataset used was from a cross-sectional observational study comparing older-aged individuals to younger-aged individuals. The demographics as well as measurements and scores from a variety of tests are included. Only patients over 60 were included for the analysis in this report because all individuals over 60 had been given a clinical dementia rating which provided the researcher a metric for dementia. Clinical Dementia Rating, age, gender, and education level for each individual is included in the data. The education level variable is an ordinal variable ranging from one to five according to the following key: 1: less than high school graduate, 2: high school graduate, 3: some college, 4: college graduate, 5: beyond college (Marcus et. al.). A categorical variable was created and used in models to help determine if college attendance had a relationship with having non-zero CDR. Thus, education levels 1 and 2 were combined and assigned "No," and education levels 3, 4, and 5 were combined and assigned "Yes."

2.1 Data Visualization

The following plot shows the quantity of individuals in each CDR level. Almost half of the individuals in the dataset have no CDR and the rest have a nonzero CDR. As expected, having no dementia is common and severe dementia uncommon.

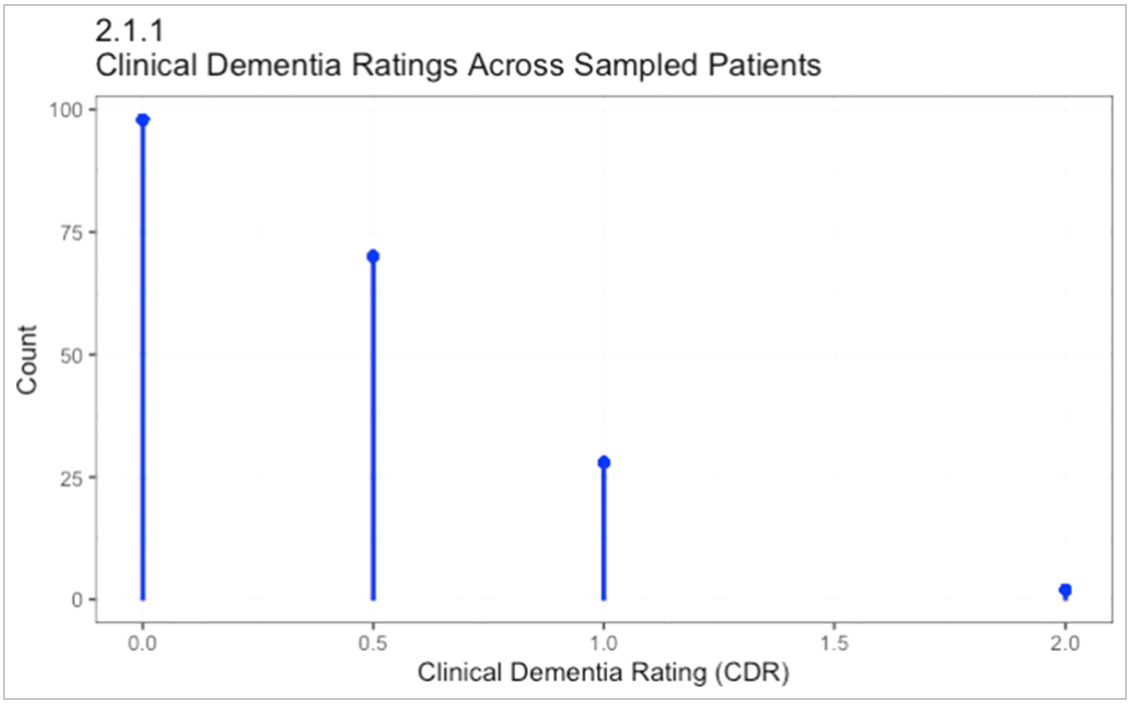


Figure 2.1.2 shows datapoints grouped by gender and college attendance and then colored by age.

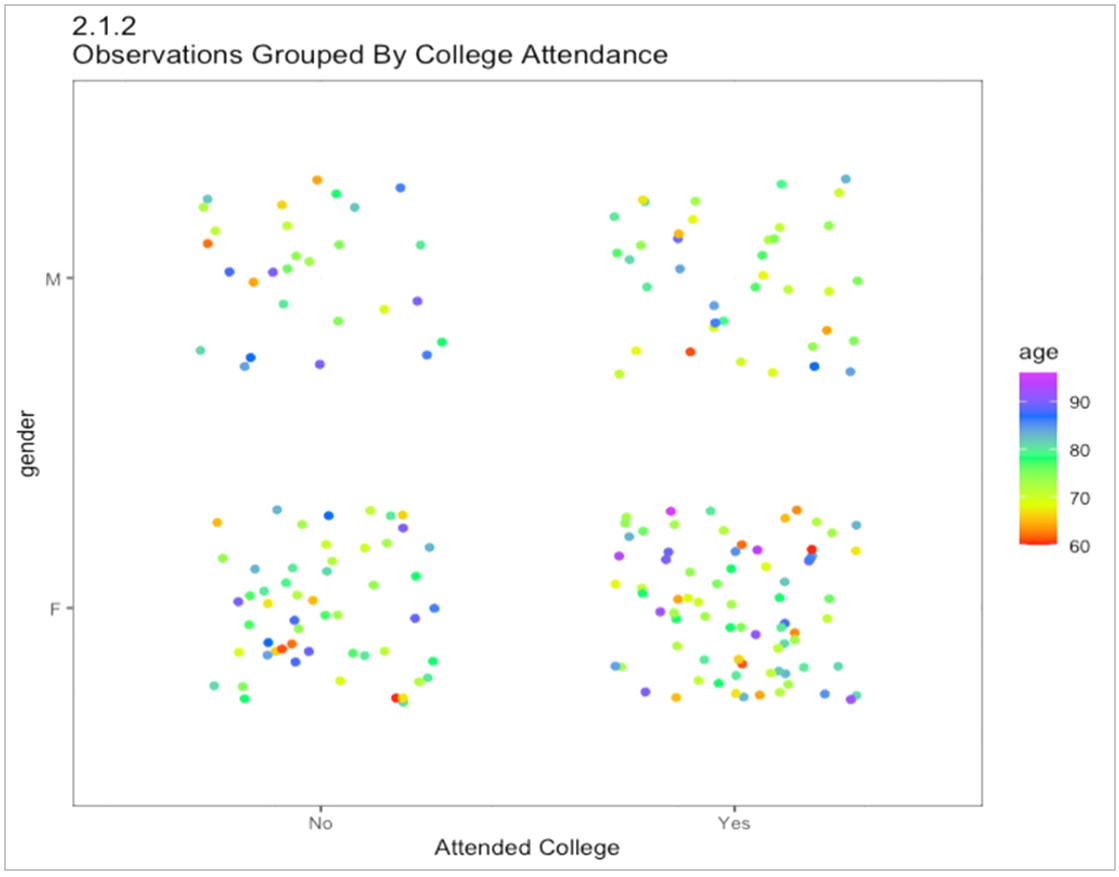


Table 2.1.3 also provides summary statistics of the different variables used in the analysis.

```
## Table 2.1.3

##     response       clin_dem_rat         age           gender
##  Min.   :0.0000   Min.   :0.0000   Min.   :60.00   Length:198
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:71.00   Class :character
##  Median :1.0000   Median :0.5000   Median :75.50   Mode  :character
##  Mean   :0.5051   Mean   :0.3384   Mean   :76.34
##  3rd Qu.:1.0000   3rd Qu.:0.5000   3rd Qu.:82.00
##  Max.   :1.0000   Max.   :2.0000   Max.   :96.00
##    education    college
##  Min.   :1.00   No : 83
##  1st Qu.:2.00   Yes:115
##  Median :3.00
##  Mean   :3.04
##  3rd Qu.:4.00
##  Max.   :5.00
```

In the end, the response variable created was a 2-level categorical variable indicating whether or not an individual had non-zero CDR. The explanatory variables were a 2-level categorical variable for gender, a 2-level categorical variable for college attendance, and a continuous variable for age.

3. Methodology

A generalized linear fit was performed, and the Logit function was appropriate, for in this case, the binary outcome was the presence of a non-zero Clinical Dementia Rating. The Logit function transforms the explanatory variables ('age,' 'gender,' and 'college') in order to relate the response variable (CDR) to $log(\frac{\pi_i}{1-\pi_i})$ where π_i is the probability of a positive outcome. This allows for the interpretation of coefficients as influencing the log-odds of the response. It is also possible to exponentiate the coefficients to interpret how the variables affect the odds of having a non-zero CDR. The research question can also be structured in statistical terms to test for any non-zero relationship between each explanatory variable and the response. Specifically, for each explanatory variable's coefficient β_i, the following hypothesis tests are relevant:

$H_0 : \beta_i = 0$ (null hypothesis)

$H_A: \beta_i \neq 0$ (Alternative Hypothesis)

The results of these hypothesis tests will be shown with the final model coefficients.

This data model came with several assumptions that needed to be met; independent, non-multicollinear explanatory variables needed to have a linear relationship with the logit of the binary response variable (Zach, B.).

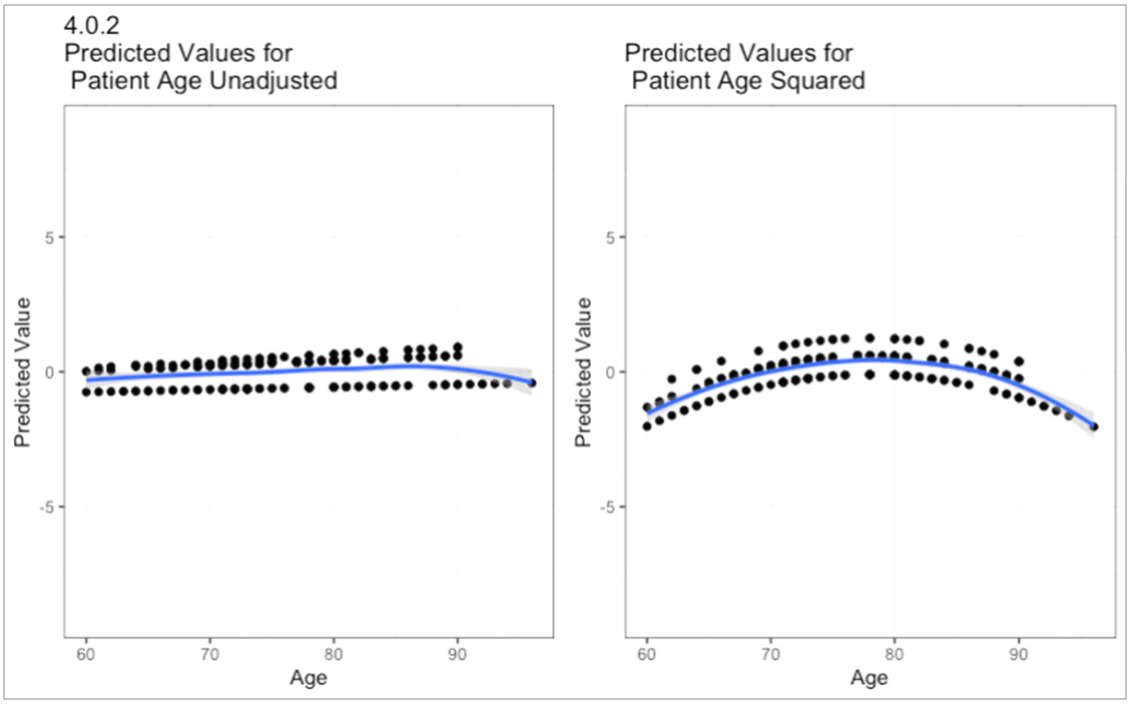
4. Results

The first model explored included all three variables with two-way interactions.

```
## Table 4.0.1

## $coefficients
##                          Estimate Std. Error     z value  Pr(>|z|)
## (Intercept)        -1.148941396 2.42194443 -0.47438801 0.6352232
## age                 0.019436620 0.03166004  0.61391652 0.5392705
## collegeYes         -0.160558580 2.87802559 -0.05578775 0.9555109
## genderM            -0.239565210 3.16943690 -0.07558605 0.9397484
## age:collegeYes     -0.010144157 0.03739458 -0.27127344 0.7861807
## age:genderM         0.006096424 0.04080476  0.14940473 0.8812343
## collegeYes:genderM  0.759566527 0.63487562  1.19640210 0.2315397
```

Not much of the variation was explained by the variables in the models, so different setups were explored. By plotting each variable's residuals, age was observed to have a squared relationship with the Logit of the response. The following two plots show predictions before and after a square transformation of the Age variable.



4.1 Final Model

After experimenting and exploring several models, the following model was selected:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 gender + \beta_4 College,$$
$$\text{where } p_i = Prob(Response_i = 1 | Age, Age^2, gender, College).$$

Table 4.1.1 shows that this model is sufficiently describing the data.

```
## Table 4.1.1

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: response
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev
## NULL                       197     274.47
## age       1   0.5394       196     273.93
## I(age^2)  1   8.7196       195     265.21
## gender    1   3.9768       194     261.23
## college   1   5.5748       193     255.66
```

The coefficients shown in Table 4.1.2 are also evidence of the effectiveness of this model as all p-values are less than our established α-level of $\alpha = 0.05$. Of particular interest is how the inclusion of the age^2 term significantly affects the explanatory power of the model. Age and gender increase the odds of having a non-zero CDR while age^2 and attending college decrease the odds of having a non-zero CDR.

```
# Table 4.1.2

## $coefficients
##                  Estimate   Std. Error   z value    Pr(>|z|)
## (Intercept) -35.604204152 13.030099074 -2.732458 0.006286360
## age           0.928960068  0.341105330  2.723382 0.006461732
## I(age^2)     -0.005957228  0.002215119 -2.689349 0.007159146
## genderM       0.637418796  0.316518196  2.013846 0.044025740
## collegeYes   -0.710276397  0.303783754 -2.338099 0.019382129
```

To satisfy the assumption of non-multicollinear explanatory variables, variance inflation factors were calculated to measure the severity of multicollinearity (correlation) in the model. This test revealed that there was no multicollinearity, so the assumption of non-correlated explanatory variables is met. It was also known from the beginning that each observation was independent, so the only assumption in need of further consideration is the assumption of a linear relationship between the transformed explanatory variables and the response. The model structure above and p-values support this claim.

5. Conclusion

From the final model's coefficients, it is possible to determine the increase or decrease in odds of having a non-zero Clinical Dementia Rating by performing a back transformation. Notably, holding all else constant, as an individual's age increases by 1 year, the odds of having a non-zero CDR increase by $e^{0.93} - e^{0.01} = 1.5$ on average meaning they are 1.5 times more likely to have a non-zero CDR. In addition, holding all else constant, on average, males over 60 years old are $e^{0.64} = 1.9$ times more likely to have a non-zero CDR when compared to similar females. Lastly, holding all else constant, for individuals over 60 years, those who attended college were $e^{-0.71} = 0.49$ times as likely to have a non-zero CDR meaning that on average, those who attended college were roughly half as likely to have a non-zero CDR.

The final model helps answer the question of which patient demographics influenced the Clinical Dementia Rating measure. More data increases the power of this analysis. To improve future analyses, more should be learned about the data collection method of the OASIS Project to ensure that this data source is reliable and that interpretations are foundationally sound.

To extend the scope of analysis, other valid measures of dementia such as Normalized Whole Brain Volume could be included (Whitwell, J., et al.). Further research could be done using the MR images themselves along with classification machine learning models to predict Dementia and even Alzheimer's Disease.

References

Boysen, J. (2017, August 16). MRI and Alzheimers. Retrieved November 15, 2020, from https://www.kaggle.com/jboysen/mri-and-alzheimers

CDC. (2019, April 05). What Is Dementia? Retrieved November 15, 2020, from https://www.cdc.gov/aging/dementia/index.html

Charpentier, A. (2013, August 23). Residuals from a logistic regression. Retrieved November 15, 2020, from https://www.r-bloggers.com/2013/08/residuals-from-a-logistic-regression/

Fox, J. and Weisberg, S. (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Marcus et. al. (2007, August 20). OASIS Brains. Retrieved from https://www.oasis-brains.org/files/oasis_cross-sectional_facts.pdf

OASIS Brains. (n.d.). Retrieved November 15, 2020, from https://www.oasis-brains.org/

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rockwood, K., Strang, D., MacKnight, C., Downer, R., & Morris, J. C. (2000). Interrater reliability of the Clinical Dementia Rating in a multicenter trial. Journal of the American Geriatrics Society, 48(5), 558–559. https://doi.org/10.1111/j.1532-5415.2000.tb05004.x

Whitwell, J., Crum, W., Watt, H., & Fox, N. (2001, September 01). Normalization of Cerebral Volumes by Use of Intracranial Volume: Implications for Longitudinal Quantitative MR Imaging. Retrieved October 24, 2020, from http://www.ajnr.org/content/22/8/1483

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Zach, B. (2020, October 13). The 6 Assumptions of Logistic Regression (With Examples). Retrieved November 15, 2020, from https://www.statology.org/assumptions-of-logistic-regression/

R Code

```
library(tidyverse)
library(car)
require(gridExtra)

data <- read_csv("oasis_cross-sectional.csv")

df <- tibble(
  clin_dem_rat = data$CDR,
  age = data$Age,
  gender = data$`M/F`,
  education = data$Educ
)
df <- df[df$age >= 60,]

college <- c()
response <- c()

for (i in 1:nrow(df)) {
    if (df$clin_dem_rat[i] != 0) {response[i] = 1}
      else {response[i] = 0}

    if (df$education[i] <= 2) {college[i] = "No"}
      else {college[i] = "Yes"}
}

df <- cbind(response, df, college)
ggplot(data = df) +
  geom_bar(aes(x = clin_dem_rat), color = "blue", fill = "blue", width = 0.005) +
  geom_point(x = 0, y = nrow(df[df$clin_dem_rat == 0,]), color = "blue") +
  geom_point(x = 0.5, y = nrow(df[df$clin_dem_rat == 0.5,]), color = "blue") +
  geom_point(x = 1, y = nrow(df[df$clin_dem_rat == 1,]), color = "blue") +
  geom_point(x = 2, y = nrow(df[df$clin_dem_rat == 2,]), color = "blue") +
  ggtitle("Clinical Dementia Ratings Across Sampled Patients") +
  xlab("Clinical Dementia Rating (CDR)") +
  ylab("Count") +
  theme_bw()
set.seed(11112011)

ggplot(data = df, aes(x = college, y = gender, color = age)) +
  geom_jitter(width = .3, height = .3) +
  scale_color_gradientn(colors = rainbow(5)) +
  xlab("Attended College") +
  ylab("gender") +
  ggtitle("Observations Grouped By College Attendance") +
  theme_bw()


college <- c()
for (i in 1:nrow(df)) {
    if (df$education[i] <= 2) {college[i] = 0}
      else {college[i] = 1}
}

df$response <- as.factor(df$response)
df$gender <- as.factor(df$gender)
df$education <- as.factor(df$education)
```

```
df$college <- as.factor(df$college)
model_init <- glm(response ~ (age + college + gender)^2, data = df, family = binomial(link =
"logit"))
model_final <- glm(response ~ age + I(age^2) + gender + college, data = df, family =
binomial(link = "logit"))
vif(model_final)
summary(model_init)[12]

p1 <- ggplot(data = df, aes(x = age, y = predict(model_init))) +
  geom_point() +
  geom_smooth(method = "loess") +
  ylim(c(-9,9)) +
  xlab("Age") +
  ylab("Predicted Value") +
  ggtitle("Predicted Values for\n Patient Age Unadjusted") +
  theme_bw()

p2 <- ggplot(data = df, aes(x = age, y = predict(model_final))) +
  geom_point() +
  geom_smooth(method = "loess") +
  ylim(c(-9,9)) +
  xlab("Age") +
  ylab("Predicted Value") +
  ggtitle("Predicted Values for\n Patient Age Squared") +
  theme_bw()

grid.arrange(p1, p2, ncol=2)
summary(model_final)[12]
knitr::kable(df[1:7,], col.names = c("Response", "CDR", "Age", "gender", "Education Level",
"College"))
```

Dataset Excerpt

Response	CDR	Age	gender	Education Level	College
0	0.0	74	F	2	No
1	0.5	73	F	4	Yes
0	0.0	74	M	5	Yes
0	0.0	81	F	5	Yes
1	0.5	76	M	2	No
1	0.5	82	M	2	No
0	0.0	89	F	5	Yes
…	…	…	…	…	…