First Year Data Analysis Qualifying Exam
Department of Statistics, University of California, Irvine

Handed out: **Monday, June 21, 2021** Due: **Friday, June 25, 2021 at 5:00pm**

Turning In Your Exam: Upload your complete solution using the link provided to you by Kazumo Washizuka (kwashizu@uci.edu) by 5pm on Friday, June 25. LATE EXAMS WILL NOT BE ACCEPTED AND WILL NOT BE SCORED.

1 Background

Alzheimer's disease is a progressive disease that destroys memory and other important mental functions. On June 7, 2021 the US FDA granted accelerated approval for the drug aducanumab (brand name Aduhelm) for reducing the level of amyloid beta (Abeta) plaques in the brain of participants at increased risk for Alzheimer's disease. These plaques accumulate in the brain and are a hallmark of Alzheimer's disease. There has been much controversy over the approval of aducanumab because the clinical trials only convincingly demonstrated that the drug reduced Abeta in the brain, but did not demonstrate clinical improvement in the cognition or functioning of participants. The FDA's approval was guided by the conjecture that reduction in Abeta due to aducanumab would ultimately slow cognitive and functional decline. In this problem we will use observational data (not a part of the clinical trials described above) to assess the relationship between the Abeta levels of participants with mild cognitive impairment (MCI) and their cognitive performance on a validated neuropsychological test called the cognitive dementia rating (CDR) scale. MCI is an early stage of memory loss or other cognitive ability loss, and can be a precursor of Alzheimer's disease.

Data available are observational in nature and follow a cohort of $N = 1,065$ individuals who entered the study and had a MCI diagnosis of early MCI (EMCI=less severe) or late MCI (LMCI=more severe). At their baseline visit, participants were offered, but not required, to undergo a lumbar puncture to measure the amount of Abeta in their cerebral spinal fluid (CSF). CSF-derived Abeta is inversely correlated with Abeta plaques in the brain (low values in CSF generally mean high values in the brain). A lumbar puncture involves inserting a needle into the lower back and extracting spinal fluid. Because of the invasiveness of the procedure many participants chose not to have Abeta levels measured.

At baseline, other participant characteristics were recorded (e.g. age, gender, years of education, race/ethnicity, APOE genetic status) along with neuropsychological test scores (e.g. CDR as previously mentioned, the Alzheimer's Disease Assessment Scale (ADAS), and the Mini-Mental Status Exam (MMSE)). Neuropsychological test results were not provided back to the participant until a week after the visit was completed. Participants also had a MRI scan of the brain to assess total brain volume as well as in two other portions of the brain associated with memory and cognition (the hippocampus and the entorhinal cortex). Participants were then followed annually and these measures were repeated over time (with exception of the lumbar puncture). Participants have data for up to 7 annual visits (84 months) after the baseline visit, for up to 8 repeated measurements.

2 Available Data

The data from the study can be found at `https://www.ics.uci.edu/~dgillen/FYE2021_DA/Abeta_MCI.csv`. Below is a brief description of the variables you have access to. Some background is given on the variables, but you can (and should) do independent reading to assess what role different covariates play in your analyses in order to address the scientific questions coming later.

- `PTID` = unique participant ID
- `MONTH` = number of months since baseline visit (baseline visit is month 0)
- `DX_BL` = baseline diagnosis (EMCI=early MCI; LMCI=late MCI)
- `AGE` = age at baseline visit in years
- `EDUCAT` = years of education
- `ETHNICITY` = self-reported ethnicity (Hispanic vs. Non-Hispanic)
- `RACE` = self-reported race (Am Indian/Alaskan, Asian, Black, Hawaiian/Other PI, More than one, Unknown, White)
- `SEX` = sex of the participant (Female, Male)
- `MARRY` = marital status of the participant (Divorced, Married, Never married, Unknown, Widowed)
- `APOE` = APOE genetic status (0, 1, or 2 copies of the e4 allele; known to be associated with AD risk)
- `ABETA_BL` = amount of Abeta measured in CSF at baseline visit
- `TAU_BL` = amount of tau protein measured in CSF at baseline visit (total tau, a component of this protein known as phosphorylated tau is known to be more specific to AD)
- `PTAU_BL` = amount of phosphorylated tau protein measured in CSF at baseline visit (like Abeta, accumulation in the brain is a hallmark of AD; highly correlated with total tau)
- `CDR` = cognitive dementia rating scale (Range: 0-18; high values are indicative of dementia)
- `ADAS11` = Alzheimer's disease assessment score (Range: 1-74; high values are indicative of less disease)
- `MMSE` = mini-mental status exam (Range: 0-30; high values are indicative of less disease)
- `HIPPOCAMPUS` = volume of hippocampus as measured by MRI
- `ENTORHINAL` = volume of entorhinal cortex as measured by MRI
- `WHOLEBRAIN` = total brain volume as measured by MRI

The first 6 lines of the dataset are below:

```
> head( abeta_mci )
    ptid month dx_bl  age educat       ethnicity  race    sex   marry apoe4
1 pt_729     0  LMCI 65.1     16 Not Hisp/Latino White Female Married     1
2 pt_729    12  LMCI 65.1     16 Not Hisp/Latino White Female Married     1
3 pt_729    24  LMCI 65.1     16 Not Hisp/Latino White Female Married     1
4 pt_729    36  LMCI 65.1     16 Not Hisp/Latino White Female Married     1
5 pt_729    48  LMCI 65.1     16 Not Hisp/Latino White Female Married     1
6 pt_729    60  LMCI 65.1     16 Not Hisp/Latino White Female Married     1
  abeta_bl tau_bl ptau_bl cdr adas11 mmse hippocampus entorhinal wholebrain
1    957.8  316.9   31.69 0.5   6.67   27        6107       2511     927913
2    957.8  316.9   31.69 3.0  11.67   23        5580       2278     915859
3    957.8  316.9   31.69 6.0   5.67   23          NA         NA     924923
4    957.8  316.9   31.69 2.0  12.00   23        5585       2241     919739
5    957.8  316.9   31.69 3.5  17.00   24          NA         NA     862436
6    957.8  316.9   31.69 5.0  21.00   26        5272       1866     867373
```

3 Scientific Goals

As previously noted, our goal is to assess the relationship between Abeta levels and cognitive performance on a validated neuropsychological test called the cognitive dementia rating (CDR) scale. As such, you are asked to use the available data to address the following aims that are broken up into two large parts:

PART 1: You will find that many participants chose not to undergo a lumbar puncture. For this part of the analysis you should subset the data and only analyze those that have a baseline Abeta value, implying that they underwent the procedure. We will consider those patients that chose not to have the procedure in Part 2.

1. Among MCI participants overall, quantify the association between baseline CSF-derived Abeta and baseline CDR score. Does the association differ by early vs. late stage MCI?
2. Among MCI participants overall, quantify the association between baseline Abeta and the trajectory of CDR scores over time. Does the association differ by early vs. late stage MCI?

PART 2: As previously noted, many participants chose not to undergo a lumbar puncture. As such, your results from Part 1 may be skewed towards the associations observed in individuals willing to undergo the procedure. For example, most of the participants will know their APOE genetic status when entering the study and those that know their APOE genetic status is positive may be more likely to undergo a lumbar puncture to learn their results. The second part of this problem will begin to assess if the association between Abeta and cognition differs depending upon a participant's decision to undergo a lumbar puncture.

3. Build a model to predict a participant's willingness to undergo lumbar puncture and identify those factors most likely to be associated with this decision. You may assume that if a participant has a missing baseline Abeta value, they chose not to have a lumbar puncture. In considering what predictors to potentially include in your model you should consider a participant's likely knowledge of the predictor value when making the choice to undergo the procedure, since you are considering factors that influence their decision.
4. Utilizing the predictions from (3), assess whether the associations you estimated in (1) and (2) vary depending upon a participant's likelihood to undergo a lumbar puncture. In particular, you may want to explore approaches that could reduce the impacts of the selection bias on your results for Part 1. In the Discussion section of your report you should discuss the relevance of any findings with respect to addressing the questions in Part 1 and highlight potential limitations of your approach.

4 General Instructions

You are to analyze the data to best address the scientific goals stated above. You should use appropriate and reasonably efficient statistical methods for estimating and quantifying uncertainty in associations. The most appropriate methods may sometime slightly extend beyond those discussed in class and require independent exploration of the literature, but you should understand and defend the appropriateness of your analytic choices in your report. Your final analysis should be presented in the form of a brief report (no more than 10 pages including primary tables and figures). You may place additional information (eg. relevant diagnostic plots) in an Appendix if you feel it necessary. The report should (at minimum) consist of the following sections:

1. Abstract - A brief summary of your basic findings.
2. Introduction - Background on the scientific problem, an introduction to the problem at hand, and what is to be addressed.

3. Statistical Methods - A clear discussion and justification of the methods you have used to analyze the data and the modeling strategy that you employed.

4. Results - A presentation of the results of your analysis that includes relevant and properly formatted tables and figures as well as complete and precise interpretations of your analytic findings.

5. Discussion - A synopsis of your findings, what you have achieved with respect to the scientific goals, any limitations your analysis may suffer from, and possible future directions to better achieve the scientific goals you set out to accomplish.

Your report should be well-written, succinct, and to the point! It should be written in a language that is understandable to the broad scientific community while precisely interpreting your finding. The discussion of statistical methods should be more technical than that provided to a non-statistical audience given the purpose of the report. It should be complete but brief - free of garbage and not-so-relevant material. It is critical that the appropriateness of your modeling choices be clearly justified in your report. You are encouraged to use relevant and well-formatted tables, plots and figures to help explain your findings. You may use any written references for this problem that you wish, **but you cannot communicate (talk, email, etc) with anyone about your analysis**.