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

Handed out: **Monday, June 25, 2018 at 12:00pm** Due: **Friday, June 29, 2018 at 4:00pm**

Turning In Your Exam: Email your complete solution to BOTH Rosemary Busta (rbusta@ics.uci.edu) and Dan Gillen (dgillen@uci.edu) by 4pm on Friday, June 29. LATE EXAMS WILL NOT BE ACCEPTED AND WILL NOT BE SCORED.

1 Background

Alzheimers disease (AD) is a type of dementia which is typically only diagnosed definitively at death upon identification of the characteristic plaques and tangles found in the brain. Consequently, diagnosis and treatment of the disease in the living relies heavily upon neuropsychological testing, imaging, and subjective clinical judgment including established guidelines for diagnosis decided by large interdisciplinary research groups and professional organizations. Built into these guidelines are a pattern of deficits that include dysfunction in certain cognitive areas which must be documented by accepted neuropsychological testing as well as documented decline over time. However, a typical problem of longitudinal cognitive testing is that repeated testing may yield increases in score upon subsequent testing times. These increases over time due to repeated testing are known as *retest effects* and may mask the very effect that is supposed to provide evidence of illness. Efforts have been made, though not with regularity, to find ways to either quantify retest effects or alleviate them, usually by using alternate test forms. Given that neuropsychological testing comprises such a large component of AD diagnosis, and that retest effects are a documented problem in repeated neuropsychological testing, clinicians would benefit from a more accurate characterization of retest effects. Quantification of retest effects may also improve diagnosis, or at least provide clarity, in ambiguous or borderline cases where scores are falling in between diagnostic categories. To better characterize neuropsychological data being collected across the nation for patients with normal cognition, mild cognitive impairment (MCI) and AD, there is interest in using data from a large national cohort study to quantify retest effects in the neuropsychological data used for diagnosis and study of disease progression.

Longitudinal data have been collected on N=15,665 unique subjects, with each subject having at least two cognitive tests and a maximum of 12. Cognitive testing visits were scheduled to occur approximately once a year, however there is variation in the times at which subjects returned for visits. In some cases, patients missed a visit and returned the following year. This is indicated by their visit numbers. Your analysis will focus on three cognitive tests: Logical Memory, Trails Making (version B), and Boston Naming. The Logical Memory test is a subset of the Wechsler Memory Scale, with total scores ranging from 0 to 25. The Trails Making test results in the total number of seconds a patient requires to sequentially draw connecting lines between alternating numbers and letters (A, 1, B, 2, C, 3, etc), and has a range up to 300. Finally, in the Boston Naming test patients are shown 30 line drawings of objects and asked to name each object before moving on to the next. Scores range from 0 to 30.

For each of the above tests, the items/questions within each test do not change from one visit to the next. Because of this it is interest to determine of retest effects in which patients scores may actually increase with increased number of visits (despite increasing age between visits). Specifically, it is of interest to determine if retest effects exist and are different depending upon whether a subject has been diagnosed as clinically normal, MCI, or demented.

Other covariates available in your dataset which may be related to test performance include basic demographics (such as age, sex, race, education) and co-morbidities such as diabetes, history of stroke, seizures, cardiovascular disease, etc.

2 Scientific Goals

The primary scientific questions you are to address are as follows:

1. Focusing **only on the difference in test scores between the first two visits** for each patient (ie. ignoring any other longitudinal followup data), determine if there are retest effects for each of the tests of interest and whether these retest effects differ by diagnosis group (normal, MCI, demented).
2. Using **all available longitudinal data**, quantify the rate of change in scores for each of the tests of interest and determine if
   (a) the within-subject rate of change over time in each test differs by diagnosis group (normal, MCI, demented),
   (b) there are retest effects present in any of the diagnosis groups.
3. Based upon your findings, suggest possible approaches to adjusting patient test scores that may account for possible retest effects (if you found any) and discuss the limitations that might be involved with your approach.

3 General Instructions

You are to analyze the data (see below) to best address the scientific goals stated above. You should use appropriate statistical methods for estimating and quantifying uncertainty in associations. Your final analysis should be presented in the form of a brief report (no more than 10 pages including relevant 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 they 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. 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**.

4 Available Data

The data for addressing the above questions is provided in the file `LearningEffects.csv`. These data formatted in a "long" format in which the exam results for each patient visit, along with the days from the first (baseline) visit are placed on a separate row. The data are available as a `.csv` file and can be found at the following site:

`http://www.ics.uci.edu/~dgillen/FYQualExam2018/LearningEffects.csv`

A brief description of the variables in the dataset is given Table 1. The first 5 lines of the data are given below:

```
> head(LearningEffects)
          id center visit visitday age.0 female race educ dx
13296 000011   1416     1        0    62      1    1   16  2
13297 000011   1416     2      427    62      1    1   16  2
13298 000011   1416     3      778    62      1    1   16  2
13299 000011   1416     4     1204    62      1    1   16  2
15434 000067   2096     1        0    60      0    1   18  2
15435 000067   2096     2      357    60      0    1   18  2
      smk30 smk100 smkyrs packyrs quitsmk alc.use alc.freq
13296     0      0      0       0     888      NA       NA
13297     0      0      0       0     888      NA       NA
13298     0      0      0       0     888      NA       NA
13299     0      0      0       0     888      NA       NA
15434     0      0      0       0     888      NA       NA
15435     0      0      0       0     888      NA       NA
      diabetes hyperten mi chf hichol seizures tbi stroke
13296        0        0  0   0      0        0   0      0
13297        0        0  0   0      0        0   0      0
13298        0        0  0   0      0        0   0      0
13299        0        0  0   0      0        0   0      0
15434        0        0  0   0      0        0   0      0
15435        0        0  0   0      0        0   0      0
      arthritis u.incont logic.memory trails boston.name
13296        NA       NA           12     76          27
13297        NA       NA           13     72          28
13298        NA       NA           12     79          28
13299        NA       NA           11     54          28
15434        NA       NA           12    105          29
15435        NA       NA           12     94          29
```

Table 1: Variable description for the `LearningEffects.csv` dataset.

Column	Variable Name	Description
1	`id` :	Unique subject ID
2	`center` :	Study center where subject was recruited
3	`visit` :	Study visit number
4	`visitday` :	Number of days between first visit and current visit
5	`age.0` :	Subject age in years at first visit
6	`female` :	Indicator of female sex
7	`race` :	Subject race (1=White, 2=?Black or African American, 3=American Indian or Alaska Native, 4=Native Hawaiian or Pacific Islander, 5=Asian, 6=Multiracial)
8	`educ` :	Number of years of education
9	`dx` :	Clinical cognitive diagnosis (1=normal, 2=Impaired/MCI, 3=demented)
10	`smk30` :	Smoked cigarettes in last 30 days (1=yes, 0=no, 9=unknown)
11	`smk100` :	Smoked more than 100 cigarettes in life (1=yes, 0=no, 9=unknown)
12	`smkyrs` :	Total years smoked cigarettes (range: 0-87, 88=not applicable, 99=unknown)
13	`packyrs` :	Average number of packs smoked per day (0=No reported cigarette use, 1=1 cigarette to less than 1/2 pack, 2=1/2 pack to less than 1 pack, 3=1 pack to 1 1/2 packs, 4=1 1/2 packs to 2 packs, 5=More than two packs, 8=Not applicable, 9=Unknown)
14	`quitsmk` :	If the subject quit smoking, age at which he/she last smoked (range: 7-110, 888=not applicable, no significant smoking history, 999=unknown)
15	`alc.use` :	In the past three months, has the subject consumed any alcohol? (1=yes, 0=no, 9=unknown)
16	`alc.freq` :	During the past three months, how often did the subject have at least one alcoholic drink (0=Less than once a month, 1=About once a month, 2=About once a week, 3=A few times a week, 4=Daily or almost daily, 8=Not applicable, no alcohol consumption in last three months, 9=Unknown)
17	`diabetes` :	History of diabetes? (0=Absent, 1=Recent/Active, 2=Remote/Inactive, 9=Unknown)
18	`hyperten` :	History of hypertension? (0=Absent, 1=Recent/Active, 2=Remote/Inactive, 9=Unknown)
19	`mi` :	History of heart attack? (0=Absent, 1=Recent/Active, 2=Remote/Inactive, 9=Unknown)
20	`chf` :	History of congestive heart failure? (0=Absent, 1=Recent/Active, 2=Remote/Inactive, 9=Unknown)
21	`hichol` :	History of high cholesterol? (0=Absent, 1=Recent/Active, 2=Remote/Inactive, 9=Unknown)
22	`seizures` :	History of seizures? (0=Absent, 1=Recent/Active, 2=Remote/Inactive, 9=Unknown)
23	`tbi` :	History of traumatic brain injury? (0=No, 1=Yes, 9=Unknown)
24	`stroke` :	History of stroke? (0=Absent, 1=Recent/Active, 2=Remote/Inactive, 9=Unknown)
25	`arthritis` :	History of arthritis? (0=No, 1=Yes, 8=Not Assessed)
26	`u.incont` :	History of urinary incontinence? (0=No, 1=Yes, 8=Not Assessed)
27	`logic.memory` :	Logical Memory Test score (range: 0 - 25)
28	`trails` :	Trails (version B) test score (range: up to 300 seconds)
29	`boston.name` :	Boston Naming Test score (range: 0 - 30)