Data Analysis - 2020 Qualifying Exam, First Year

Department of Statistics, University of California, Irvine

Handed out: Monday, June 22 , 2020 Due: Friday, June 26, 2020 at 5:00PM

Turning In Your Exam: Email your complete report as a pdf file to BOTH Annie Qu (aqu2@uci.edu) and Kazumo Washizuka (kwashizu@uci.edu) by 5pm on Friday, June 26. LATE EXAMS WILL NOT BE ACCEPTED AND WILL NOT BE SCORED.

Background of ACTG Data Analysis: The Harvard AIDS clinical trial group (ACTG) data is a longitudinal data study to evaluate the treatment effect of Zidovudine on CD4 cell counts (e.g., Dolin et. al., 1995). CD4 levels represent a surrogate endpoint for HIV positive individuals as lower CD4 counts are associated with shorter times to progression of AIDS and survival times. You have access to a subset of data on 265 patients from the ACTG randomized clinical trial comparing Zidovudine to control. All individuals included in this subset had CD4 counts above 50 at baseline (time of randomization) and longitudinal data on CD4 counts were assessed at up to 14 time points. Measurements were taken at approximately 1 month intervals. Patients might have intermittent missing or dropout with an overall missing rate of 20.2%.

In the dataset, we denote $Trt = 1$ if the patient receives the treatment Zidovudine and $Trt = 0$ if the patient is in the control group. $Time$ represents the measurement number for CD4 ($Time = 1$ is baseline). Age is measured in years, and $Gender = M/F$ for male/female. The primary scientific aim of your analysis is to quantify the effect of treatment with Zidovudine on the trajectory of CD4 counts over time.

Objectives of Data Analysis:

The overarching goal of your analysis is to quantify the effect of treatment with Zidovudine on the trajectory of CD4 counts over time using the provided data (`ACTG.csv`) (`NA` means missing response). In accomplishing this goal you are asked to incorporate the following specific components into your solution. While the below components are somewhat directed, your final solution should include a complete scientific analysis including informative descriptive statistics, data visualization, model diagnostics, etc for answering the scientific questions of interest. Further details on the report write-up are given later.

Specific components that should be included in your analysis:

(a) Use generalized estimating equations (GEE) with an appropriately chosen working correlation structure to quantify the effect of treatment with Zidovudine on the trajectory of CD4 counts over time.

(b) Re-analyze the data using a linear mixed-effects (LME) model with appropriately chosen variance components.

(c) Using your models in (a) and (b), provide estimated trajectories of CD4 counts for 4 randomly sampled patients (2 sampled from the Zidovudine group and 2 from the control group) and compare them with their observed counts. For these 4 patients, also provide predictions of their CD4 counts at 1 month and 3 months after that last known measurement. Compare and contrast the GEE and LME estimates.

(d) Explain different interpretations corresponding to the GEE model and the linear random-effects model based on this data. In which scenario, might one prefer one modeling approach over the other?

(e) Specifically comment on missing data in your analysis and any potential impact it may have on your results and findings.

General Instructions

You are to analyze the data to best address the scientific question of interest. You should properly justify your model and use appropriate statistical methods for estimating and quantifying uncertainty in associations. Your solution should also include responses to each of the above components.

Your final analysis should be presented in the form of a brief report (no more than 10 pages including relevant tables and figures). A font size of 11 points or larger must be used. Margins in all directions must be at least one inch. You may place additional information (eg. diagnostic plots) in an appendix if you feel it necessary; however, the appendix should be readable: do not copy-paste your computer code and output. The report should (at minimum) consist of the following sections:

1. Abstract - A brief summary of your basic findings

2. Introduction - A brief introduction/motivation to the problem at hand and what is to be addressed

3. Statistical Methods - A discussion and justification of the methods you have used to analyze the data

4. Results - Present main conclusions of your analyses.

5. Discussion - A synopsis of your findings and any limitations your study may suffer from.

Your report should be succinct and to the point. It should be written in a language that is understandable to the scientific community. You may use tables, plots and figures to help explain your findings. You may use any written references for this problem that you wish.

You cannot talk to anyone about your analysis. If you need clarification about anything you may ask Annie, but only before June 25. Please type at the beginning of your exam: **I spoke with no-on concerning this exam except for a faculty member** and then type the last four digits of your student ID.

References:

Dolin, R., Amato, D. A., Fischl, M. A. et al. (1995). Zidovudine compared with Didanosine in patients with advanced HIV type 1 infection and little or no previous experience with Zidovudine. *Archives of Internal Medicine,* ***155****, 961-74.*