Project I, BIOTAT 620W2024

Due: Friday, February 23, 2024

Do this project with a group and submit the project to Canvas. AI tools (e.g. ChatGPT or as such) are not allowed to be used for this project.

The objective of this project is twofold:

(i) To understand and practice the operation of federated learning techniques and distributed computing as a primary approach to protecting data security and privacy while analyze the screen use activity data collected from a group of 2-3 people. Set **the data freeze date as of February 13, 2024**.

(ii) To organize and create a combined group-level data sheet that will be later shared with the entire class in Project II. This data sheet will assemble updated data from all individuals in a group, containing the screen activity data and subject-level baseline covariates. **This data sheet must be updated periodically (say, weekly) through the entire winter semester.**

(iii) Set up a GitHub of your group to save R functions, descripts and demos of Project I.

Organization of the project.

Title: Give a title of project I.

Author: List all authors who make primary contributions to the project.

Abstract: Write a high-level summary of project I.

Key Phases: List five key words or less that allow people to search your project for relevant content. These key words should have minimal repetition with the project title.

Introduction: This section should include objective, hypothesis, and motivation of your project, followed by a brief discussion why your data are relevant to your proposed study. This section needs to present a summary of your data analysis as well as major findings. In addition, this section should include some discussions about significance of your study, innovation of your study as well as some relevant references to support your proposed study aims. Adding an relevant figure for the background science, organization of your study, or schematic procedure of your algorithms may make your project more interesting or attractive to the audience.

Data Description: This section covers background of data collection, list of variables (see a separate file), including personal variables and team features. It is important to create the so-called "Table 1'' that lists summary descriptive statistics for all collected variables to describe the team. You may also consider using figures (e.g. boxplot, histogram, scatterplot, time series plot, ACF plot, occupation time curve, and circular plot of 24 hours), if necessary, to display patterns and distributions of some variables for which you like to show more detailed information.

Data Preprocessing: This section discusses issues related to data merging from individual data sheets into a group data sheet, data cleaning and validation, data harmonization in data merging, variable transformation and so on.

Federated Learning: Based on your study objective and hypothesis given in the Introduction

section as well as preliminary data analysis in the Data Description section, in this section you choose an outcome variable y (e.g. daily social screen time), and a set of predictors $x_1,\ldots,x_p$ (e.g. endogenous variables: daily number of pickups, daily first pickup time, etc; exogeneous variables: weekday/weekend, before/after the winter semester began, snowy day yes/no, etc.), to run a linear regression $y \sim x_1 + \ldots + x_p$ via the federated learning method. **Here, each device user represents a data source where raw data cannot be shared, but summary statistics are allowed to be shared.** (a) First, describe a federated learning procedure, including model assumptions, for the calculation of regression parameters (intercept, slopes and variance) and their standard errors as well as goodness-of-fit. Second, (b) design a distributed computing platform to implement your federated learning machinery developed in part (a). (c) Third, report your analysis results in forms of figures and/or tables.

Confirmation analysis: Note that your team can create a combined data sheet of the raw data. Thus, you can repeat the above regression analyses *using the combined raw data to obtain the oracle results*. Compare and confirm numerically if the results from the federated learning method above are the same as the oracle results. Use residual analyses to check the model assumptions that cannot be done in the federated learning paradigm.

Conclusion & Discussion: Summarize your main contributions and findings in this project, including comments on heterogeneity assumption for the regression parameters. What was your experience of data analysis, especially in the aspect of team collaboration? What have you found to be most interesting and surprising? What are the limitations of your study (e.g. the inclusion of confounding factors in the analysis)? What is the future work (e.g. seemingly unrelated regression for multiple outcomes such as daily total social screen time and daily number of pickups)?

Acknowledgement: You may write some additional notes related to help given by people outside of the authorship and roles that individual authors played in the project.

References: List the literatures cited in the project.

Appendix: You can always create an appendix to include more detailed supplementary information of your project if necessary.

Format Required by the Project

Each project should be prepared with one-inch margins, in 12-point size letters and no more than 25 lines per page, double-spaced throughout. The first page should include a title, authorship, key phrases, and a one-paragraph abstract. The abstract should not exceed 200 words. **Each project should not exceed 10 pages**, including title, authors, abstract, key phrases, figures and tables as well as references. You are allowed to submit an appendix that contains some technical details, R scripts, and additional analysis results. **The appendix should not exceed 5 pages. All R scripts should be saved in your group GitHub whose link is provided.**