BIOSTAT 651, Winter 2022
End-term Project

Veera Baladandayuthapani & Kevin He

Last updated: March 14, 2022

Project Specifics

- Goal: the overarching goal of the project is the apply and evaluate the data analytic techniques learnt in this course to real-world datasets.
- You have a choice of 3 datasets, from which you can choose one for your project. It is expected that you will use some of the GLM-based models that have been covered in this course.
- This is a group project with 3-4 members in each group. Your group assignments can be accessed here:
  - Section 1
  - Section 2
- There are two components to the project: 1) Project report and 2) Project Presentation
- The report will be due **April 24th, 11:59pm** and the presentations will take place **April 25th 1-5 pm in room M1020 SPH**; No exceptions to the due date will be made.
- You are encouraged to talk to the instructors or GSI in the interim during office-hours (or by appointment) regarding any questions about the project.

Project Report

- Each group should provide a three-page report (maximum; excluding references; 11 pt font, reasonable margins). You can have appendices and/or supplementary materials to attach additional results, code etc.
- Format: your report should include the following sections
  - Introduction
  - Methods
  - Results
  - Discussion
  - Contribution from each individual team member
- The report should include a brief description and objective of your project, and indicate the proposed data analytic method to accomplish the task. Clearly state in your report the models (mathematically), the outcome variable(s), the potential predictors you have studied, exploratory analyses as well as the model building and analyses process.
- Note, it as to be your own work – you cannot just reproduce any analysis already published in the literature or elsewhere, or posted on the internet.

Project Presentation

- Your group will be required to prepare a **7-minute (strict)** presentation to describe your findings. This can be individual or a group presentation. We will reserve $\sim$**1-2 minutes** for Q&A and all members of the group should participate in the question session.
- Format of presentation:
  1. Introduction of the problem and data to address this problem
  2. Methods to address this problem
  3. Results based on your methods
  4. Discussion
  5. Q&A
- Please add your link to slides **(no more than 10 slides)** in the google sheet by April 25th 12pm.

Grading

- By default, a combined score will be given to each team and individual members (counted towards a maximum of 15% of your final class grade).
- If you feel otherwise, please send the instructor an email about alternative/preferred ways of assessing individual contributions (need to be specific about who contributed to which sections).
- Your grades for the project will be a combination of 1) the final report (70% weight) and 2) the presentation (30% weight), and will be available to you prior to your final exam.
- For both components, the grading criteria will be
  - 10% Description of scientific question and statement of aims
  - 20% Description of analytic approach
  - 20% Use of data
  - 30% Appropriate use of concepts and/or methods from class
  - 10% Interpretation of results in the scientific context
  - 10% Quality of presentation (writing in the report/ delivery in the oral presentation

Peer Review Assessment

- Grade for working compatibility and integration of each of your team member:
  - 1 (Little or weak effort)
  - 2 (Insufficient effort)
  - 3 (Sufficient effort)
  - 4 (Very strong work)
  - 5 (Excellent work)
  - Additional comments are needed if you gave 1 or 2 for your team member
- Please be as honest and fair as possible in your assessment
- Interim checkpoint (due Monday, April 4th):
  - Session 1: Click [anonymous link]
  - Session 2: Click [anonymous link]
  - Will not be used for grading
- Final evaluation (due Monday, April 25th):
  - Session 1: Click [anonymous link]
  - Session 2: Click [anonymous link]
  - Will be used for grading

Dataset #1: Melanoma

- Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. Cancer is inherently a genetic disease —that is, cancer is caused by certain changes to genes that control the way our cells function, especially how they grow and divide. Thus, when investigating various clinical cancer outcomes, it is often valuable to investigate how different genes are associated with outcomes across different tumors in different patients.
  - The objective of this project is to examine the association between various clinical outcomes of interest and gene expression data for Skin Cutaneous Melanoma (SKCM) – a type of skin cancer.
  - You have been provided with four clinical outcomes of interest along with gene expression data for 24 genes across 240 patients, a subset of the total SKCM patients and gene expression data collected by the The National Cancer Institute through the TCGA project.

Dataset #2: Dementia

- According to the World Health Organization, around 55 million people currently have dementia with projections of 139 million in 2050. Dementia affects a person's cognitive function, including memory, thinking, orientation, comprehension, and judgment. This disease mostly affects the older population, but it is not a normal part of aging. We know that dementia is caused by several diseases or disorders, including Alzheimer's disease, Lewy Body dementia, and frontotemporal disorders, but it is important to assess risk factors for these diseases or disorders.
  - The objective of this project is to conduct a statistical analysis investigating the association between cognitive impairment or dementia (outcome) and self-reported depression (main variable of interest).
  - The main scientific question of interest involves clinical diagnosis of dementia and self-reported depression on 900 patients. Other variables will be important for modeling so that we can adjust for other possible factors such as age, sex, race, dementia status of relatives, marriage status, and education.

Dataset #3: Kidney Transplantation

- End-Stage Renal Disease (ESRD) is one of the most deadly and costly diseases in the US. While a kidney transplant is the preferred treatment for ESRD, the demand far exceeds the supply. Identifying risk factors associated with post-transplant mortality is pivotal in prolonging the survival of transplant patients and optimizing organ allocations.
- Post-transplant mortality outcomes are influenced by both the performance of transplant centers, donor characteristics (e.g. donor age, race and comorbidities) and transplant recipient characteristics (e.g. age, race and comorbidities).
  - The analysis cohort included 2,436 kidney transplant recipients from 19 transplant centers. Outcome is defined as 5-year post-transplant patient status (living, death, or graft failure).
  - The objective of this project is to examine risk factors associated with post-transplant mortality and graft failure.

Remarks

- Detailed descriptions, data and additional information can be found on the class canvas site (linked in the homepage).
- Please contact the instructors with any questions about the data format and structures.
- Thanks to Nicole Wakim, Nate Osher and Di Wang for their help in collating the datasets.
- Think like collaborators and contributing to team science!
- Have fun!