AMS 572 Data Analysis I
Group Project

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Group Project

- 25 groups (3 members each group)
- The instructor will make initial group assignment randomly. You are allowed to swap groups (it will be a one-to-one swap, i.e., you need to find another member in your new group to agree on swapping)
- Deadline for group swapping is October 06, 2020 at 5PM.
- Email the instructor your new group, cc'ing the student who is willing to swap with you.
- Find a dataset which has $n \geq 50$ samples and $p \geq 10$ variables
- Your dataset should have both categorical and continuous variables
- Formulate 2 hypotheses of interests. The first hypothesis can be answered using any of the inferences that we have learned in class or more advanced methods.

Group Project

- The second hypothesis will be answered using one of the following methods
  1. Multiple linear regression + effect of missing values (Chapter 11)
  2. Analysis of multi factor experiments + effect of missing values (Chapter 13)
  3. Resampling method + effect of missing values (Chapter 14.6)
  4. Mixed effects model + effect of missing values
  5. Generalized linear model + effect of missing values
- If you have other "advanced model" for the second hypothesis, please check with the instructor to get permission to fit such model.

Group Project

Investigating effect of missing values:

- Discuss the effect of missing values on the data analysis for two scenarios
  - The missing values are completely at random. You can randomly set (additional) 20% of the data to be missing and reanalyze the data. Report the effect of missing values on data analysis.
  - Non-ignorable missing values (i.e., the missing data mechanism is related to the missing values, e.g., the patients who are sicker do not come for follow up). For non-ignorable missing values, do some literature search and write a paragraph or two describing how you will handle non-ignorable missing values in your datasets.

Timeline

- Write a report (maximum 20 pages, font size ≥ 11). Please include introduction and data description.
- Include your SAS or R source code, together with the data. If the data is large, use SBU google drive to upload the data.
- The project report is due 12/01/2020 (Tuesday) at 10.00AM via email.
- One of the group member will submit the report to the instructor (peifen.kuan@stonybrook.edu) and cc the other two group members and your TA (weihao.wang@stonybrook.edu).
- In the "Subject" header of the email, type "AMS 572 Fall 2020 Group XX Project", where XX is your group number.

Some potentially useful data archives/repositories

- Data and Story Library (DASL)
  `http://lib.stat.cmu.edu/DASL/`
- NIST Statistical Reference Datasets
  `http://www.itl.nist.gov/div898/strd/`
- UCI Machine Learning Repository
  `http://archive.ics.uci.edu/ml/`
- UMASS Datasets
  `http://www.umass.edu/statdata/statdata/data/`

You may also use your own dataset.

Grading criteria (not in the order of importance)

- Is the report well-organized?
- Is supporting computer output provided (in edited form, that is, edit out all the extraneous information in the report)? R/SAS must be used for model fitting and plotting
- Is the model appropriate for the design and questions of interest? Have you checked the assumptions?
- Are correct interpretations given for the parameters in the model?
- Are conclusions drawn from the model correct and do they answer the question of interest?
- Are the hypothesis tests interesting and non trivial?
- Can the instructor reproduce the results reported using the code the group provided if the model is correct?

Group synergy

- If you have concerns with non-contributing members and are not able to resolve within the group, please speak to the instructor immediately (do not wait till project due date).
- There will be an optional peer evaluation within each group:
  - If all the group members contribute equally to the project, you do not need to fill in the within-group evaluation.
  - Otherwise, for each member of the group, fill in the peer evaluation in the scale of 0-100%, how much each of the other group member should get from the group project score (e.g., if the group gets 25/30 and member A is given 60% by member B and 80% by member C, his/her final score will be 17.5/30). The instructor will arrange for zoom meeting with these groups to ensure fair evaluation.