ISyE 6414 — Regression Analysis

Deadlines of Course Project

For the course project, you are encouraged to form a team of $2 \sim 4$ students, and it will be OK if you decided to work on your own (in the case of team, you will need to submit only one product per team).

You are encouraged to choose a project related to your own research interests, and please feel free to discuss your project with the instructor/TA in piazza if you want.

Grading: The course project will be peer graded, and will count 20% = 2% (proposal) + 8% (oral presentation) + 10% (written report) in your final course grade.

Be aware of the following requirements and deadlines (note that we moved the deadlines from weekdays on the syllabus to the weekends based on the students' feedback).

1. **The Project Proposal** (2 points) is due at Canvas **at midnight on Wednesday, Oct 27, 2021, ET** (One submission per team). Please make sure that all teammates sign up to the same group at Canvas through "People" and "Project Group"), and you need to sign up even if you are working along. See the page #3 of the pdf file for possible datasets for your project.

   The purpose of the proposal is to get you started, and also allows others to provide feedback to your project. It shall **be** $1 \sim 2$ **pages**. You will need to provide the following information:

   (a) Your name(s)
   (b) Project description
   (c) How and where you obtained the data. For the data set, you can just direct to a website where we can find them.
   (d) Scientific Research questions you may want to address and corresponding data mining & statistical learning methods

   Peer grading: when you are assigned to grade proposals, please provide comments to the proposal, (e.g., on problem formulation whether the project sounds interesting, on dataset whether the dataset can help answer the questions, on the proposed methods, etc.). Note that all teams will receive full credits (2 points) on the project proposal as long as the team provide all these information.

2. The **Oral presentation** (8 points) of your team project in the last several lectures of the class. **The oral presentation file** will be due at Canvas **at 10am on Monday, Nov 29, 2021, ET** (either pptx or pdf version will be fine). One submission per team.

   (a) In order to be fair to all teams, we discourage the teams to update the presentation files after the deadline. However, if your team or you really think you need to do so (e.g., find critical mistakes/errors), you are allowed to update your files with a penalty, and please email your updated presentation file to the instructor at (ymei@isye.gatech.edu) as an email attachment before the class of your presentation.
   (b) There are no specific guidelines on the presentations, and the commonsense applies, e.g., write all team members' names somewhere on the first slide, highlight your problem, data set, main ideas/methods, and conclusions. Ideally you or your team should prepare for the slides so that each teammate member can present about 3 minutes with a standard deviation of 0.5 minutes, e.g., a team of 2 students will prepare for slides for about 6 ± 1 minutes presentation, and a team of 3 students should prepare for slides for 9 ± 1.5-minutes presentation, etc.

   Peer grading: please assign one of the following grades, and write comments to justify your grades:

   - 8.0 points (=100%,A+) if the presentation file is clear, the project sounds interesting, and the conclusion sounds reasonable, etc.
   - 7.2 points (=90%, A) if you have some minor concerns on the presentation file
   - 6.4 points (=80%, B) if you have some major concerns on the presentation file
   - 5.6 points (=70%, C) if you think the presentation file contains some critical technique errors or has poor presentation
   - 4.8 points (=60%, D) if you think the presentation file is not understandable or sloppy
   - 0 points if no submission

3. The **final written report** (10 points) of your team project is due at Canvas **at midnight on Monday, Dec 6, 2021, ET** (One submission per team). Either word or pdf file is fine. See the page #5 of the pdf file on some suggestions on the writeup of your report.

   (a) In your writeups, we expect clear explanations of models chosen, hypotheses tested, and findings analogous to what you would produce for a consulting project.

   (b) **Mandatory subsection in the final written report: the lessons you learned** (you can use any names for this subsection). For the purpose of this class, at the end of conclusion section of your final report, please **add a subsection for lessons you learned** from this project or this course. You can also write any suggestions to the instructor. The instructor/TAs will read this subsection, so that we can improve our teaching in the future.

   Peer Grading: At the high level, the grade will depend on the team's selecting and adhering to a logical and readable format for the report; on the appropriate use of whatever data mining technique the team uses; on the appropriateness in the conclusions of the report; and on the readability and understandability of the report when technical material is needed.

   Please assign one of the following grades and also write comments to justify your grades:

   - 10 points (=100%,A+) if you think this is an outstanding project that deserves possible publication
   - 9 points (=90%, A) if you think the written report is excellent and the project is solid from the technical viewpoint
   - 8 points (=80%, B) if you have some concerns on the project or report
   - 7 points (=70%, C) if you think the project contains some critical technique errors or the report has poor presentation
   - 6 points (=60%, D) if you think the project or report is not understandable
   - 0 points if no submission
   - The Instructor/TA keeps the right to deduct 2 points if we find out that the team miss the mandatory subsection on the lessons learned.

   As always, if you or your team has a concern about peer grading, please feel free to let the instructor/TA know asap at piazza: we will read your report on our own to make a final decision, although please do understand that ultimately the grade on the final written project will be subjective.

4. **Peer evaluation form:** if there are two or more students on a team, each teammate should also independently submit the completed **peer evaluation** form at Canvas **at midnight on Monday, Dec 6, 2021, ET** (e.g., the same due date of the final written report). If you conduct the project by yourself without team, this is optional and not required.

   See the last page for the blank peer evaluation form. This peer evaluation is to discourage free ride, and allows the Instructor to adjust an individual student's score based on the teammates' peer evaluations if needed. In general, the ideal is for all team members to receive the same grade on the final project. However, individual deductions from the team's final project grade will be assessed for failing to contribute a fair and significant share to the team's project, as determined by the teammates' peer evaluation and the instructor.

Possible Topics of Your Project

The objective of a class project is to help you gain experience with research, and to relate what you learn to real life problems which may require you learn new techniques (or develop new methods by yourself). You are expected to present the project findings during the class and submit a summary report at the end of the semester. Below are two types of possible projects, and you only need to choose one of them.

1. **Solving a real life problem.** A typical report includes problem formulation, data analysis, proposed solutions, and interpretation of results. The data set can be from your own research or the public domain.
2. **Numerical study of statistical methods/models using existing data sets in the literature.** Ideally your approach is substantially different from those in the literature, but it will be all right if you repeat the analysis as long as you did independently. Some possible projects can be
   - Compare performance of competitive statistical (or data mining) techniques;
   - Ask different questions or investigate new ideas of statistical methods or models;
   - Identify optimal parameters of specific statistical methods or models;

Note that the crucial aspect of your project is **to analyze some data sets and justify your conclusions,** not using some specific statistical methods or models we discussed in class.

Datasets: You can collect the data by yourself, use the data set from your own research or the public domain. The followings are some examples of online datasets (you can use google or other search engineer to find more):

1. http://kdd.ics.uci.edu/ or http://archive.ics.uci.edu/ml/
   One example is the KDD cup 1999 data at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
   More KDD cup data can be found at http://www.sigkdd.org/kddcup/index.php
2. http://lib.stat.cmu.edu/DASL/
3. http://www.quandl.com/ (financial and economic time-series datasets)
4. https://datamarket.com/topic/list/ (a privately held Icelandic company that specialises in providing access to data from public, and, to a lesser extent, private institutions and companies.)
5. http://www.kdnuggets.com/datasets/index.html (links to more data repositories.)
6. One of the datasets in the Appendix C of our textbook, see page 1348-1357 (except Data Set C.5 Prostate Cancer, which will be analyzed in class). To inspire you how to analyze these data sets, also see the "Case Studies" of our textbook on pages 153, 342, 382, 420, 480, 508, 554, 640, 732, 774, 809, 879, 891, 950, 990, 1028.
7. You can also obtain some data sets from me and report your findings. In this case, please email me (ymei@isye.gatech.edu) to schedule an appointment so that I can explain the data set and the questions to you and your group (preferable sometime on Mondays and Wednesdays after the class).

To inspire your projects, some concrete examples can be as follows:

- analyze some data sets in some competitions, see the links
  < http://www.kaggle.com/competitions>
- model data from some government websites such as <http://www.cdc.gov/biosense/correlate/> or <http://www.ngdc.noaa.gov/stp/satellite/goes/dataaccess.html>.
- find the traffic pattern near Georgia Tech or your appartment/home by using the traffic count data from <http://www.dot.ga.gov/informationcenter/statistics/TrafficData>
- predict Allergy season by using Atlanta Pollen count data from
  <http://www.atlantaallergy.com/PollenCount.aspx> .

ISyE 6414 — Regression Analysis

Guidelines on the Final Report

In your final summary report, we expect clear explanations of models chosen, hypotheses tested, and findings analogous to what you would produce for a consulting project. The most important advice is to follow your common senses to make your final report understandable to an intelligent scientist who might not be familiar with your project.

The main body of your final summary report (e.g., without appendix and figures/tables) is generally $5 \sim 10$ pages, and the total length of the final report shall **not be longer than** 20 **pages**. Only very relevant plots and tables shall be included in the body of the report, and the rest should go to Appendix. When writing up your summary report, it is useful to ask yourself the following questions: What is the work? Why is it important? What background is needed? How will the work be presented?

Here is a suggested format for your summary report.

1. **Title Page** (cover page): Project Title, author(s) (names, the last three digits of student IDs, and email addresses), the submission date, course name/number;

2. **Abstract:** informative summary of the whole report (100-500 words).

3. **Introduction** includes problem description, motivation and challenge(s), problem solving strategies, accomplished learning from the applications and outline of the report.

4. **Problem Statement or Data Sources**: cite the data sources, and provide a simple presentation of data to help readers understand the problem or challenge(s).

5. **Proposed Methodology**: explain (and justify) your proposed methods or models.

6. **Analysis and Results**: present *key findings* when executing the proposed methods or models. For the benefit of readability, detailed results should be placed in the Appendix. Reference of computer softwares to implement your proposed methods or models (even it is a web page) should be given.

7. **Conclusions:** Draw conclusions from your data analysis practice. Unfinished or possible future work could be included (with proper explanation or justification).

   Remark:* **at the end of conclusion section of your final report, please add a mandatory subsection for lessons you learned from this project or this course. You can also write any suggestions to the instructor.**

8. **Appendix:** This section only includes needed documents to support the presentation in the report. Feel free to divide it into several subsections if necessary. Do NOT dump all computer outputs unorganized here.

9. Bibliography and Credits.

Parts 3-6 constitute the meat of the paper for your primary audience. Usually, as with fictional boss in this example, your audience is intelligent but unschooled in Statistics. So these parts should have as little technical material as you can possibly get away with.

It is appropriate, and even recommended, to refer the reader to the appendix in part 8 if you need to provide a more technical explanation for something. Part 8 is your secondary audience - me - and should follow closely enough the "story" of parts $4-6$ that it is easy for me to see what technical material backs up with results and discussion.

It is not necessary to number these parts 1-9 or name them as-above-mentioned. Please feel free to merge some parts or provide more informative section names if it seems natural to do so.

ISyE 6414 Class Project Peer Evaluation

NAME: _____ DATE: _____

A. Suppose your team is just given $1,000 per person for work performed on this project, i.e., a 2-person team receives $2,000, a 3-person team receives $3,000, etc. Redistribute this sum to the members of your team (including yourself) based on your overall assessment of the total value of work (amount and quality) each member has contributed to the project from the beginning until now.

Name	Amount

B. Provide a brief, candid explanation of your overall assessment in Part A.

The Georgia Tech Honor Code requires that you fill out this form truthfully.

Sign here to verify that your answers represent the truth, the whole truth, and nothing but the truth:

_____