ISyE 6414 **Regression Analysis** **Spring 2020**

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 report per team). To maximize the learning of all students, we generally discourage a team of 5 or more students, though the exception is possible if you can convince the Instructor that doing so is in the best interest of everyone involved.

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

Be aware of the following deadlines (**To save the trees, please submit your presentation file and final written report to Canvas** (one per team)):

1. **March 12 (Thursday)**: the Project Proposal is due. The purpose of the proposal is to get you started. It also allows the instructor to provide feedback to your project. It shall **be** $1 \sim 3$ **pages** (or possibly more if necessary). You will need to provide the following information:

   (a) Your name(s)
   (b) Project description
   (c) How and where you obtained the data
   (d) Scientific Research questions you may want to address
   (e) The proposed statistical methods and models (this can be changed later).

   You will receive full credit on the project proposal if you provide all these information. For the data set, you can just direct me to a website where I can find them.

2. **7am on April 14 (Tuesday):** the Presentation file of your course project is due at Canvas (either pptx or pdf version will be fine). Your presentation file should be named "*Team-xx.pptx*" or "*Team-xx.pdf*" (where "*xx*" is your team ID which will be assigned by the instructor when providing the feedback to the proposal).

   Late submission can only be done through emails; the subject of your email (for late submission only) should be: **6414 Presentation submission by TEAM ID.** In order to be fair to all students, especially those who present their projects early, we discourage the update of the presentation slides after the deadline. Of course, if you find some major mistakes in your presentation slides/files (e.g., some major computation errors or false conclusions), you can email me the latest corrected version, but it will be counted as a late submission and has 20% deduction penalty.

   Ideally all students are expected to present their findings during the last week of the semester, and it is expected that all team members need to present in case of a team.

3. **April 21 (Tuesday): the Final Report is due at Canvas**. Please name your report file as "*Team-xx.doc*" or "*Team-xx.pdf*." In your writeups, we expect clear explanations of models chosen, hypotheses tested, and findings analogous to what you would produce for a consulting project.

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

Grading: The course project will have a weight of 35% = 2% (proposal)+ 15% (oral presentation or slides) + 18% (written report) in your final course grade. In general, the ideal is for all team members to receive the same grade on the course project. However, individual deductions from the team's final project grade can be assessed for failing to contribute a fair and significant share to the team's project, as determined by the instructor through the team's presentation and discussions.

Your grade on the project will be very subjective, and will depend on you selecting and adhering to a logical and readable format for the report; on the appropriate use of whatever statistical methods/models you use; on the appropriateness in the conclusions of your report; and on the readability and understandability of the report when technical material is needed.

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** **Spring 2020**

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.