Statistical Modeling and Regression Analysis - ISYE 6414

Instructor: Dr. Nicoleta Serban

Head TA: Olaoluwa (Dami) Alebiosu

Course description:

An introduction to commonly used regression models, along with the implementation of these models using data examples and statistical software.

Course prerequisites:

A sound familiarity with undergraduate or graduate-level statistics and probability, as well as proficiency in programming, linear algebra, and calculus.

Recommended Textbooks:

Two primary textbooks are highly recommended:

1. Kleinbaum, D. G., Kupper, L.L., Nizam, A., Muller, K.E. (2013) *Applied Regression Analysis and Other Multivariable Methods*, Cengage Learning; 5 edition.
2. P. McCullagh, J.A. Nelder (1989), *Generalized Linear Models*, Chapman & Hall.

Note that these books are not required. They are only meant to complement the learning of the course material. You will also find additional material from alternative resources, but you need to be careful since not all of it is rigorous and consistent with the fundamentals of statistical modeling.

What will students learn on this course?

In this course, students will learn the basics of regression analysis, including linear regression, generalized linear regression, and model selection. Students will receive a fundamental grounding in the use of widely used tools, but much of the course's energy is focused on individual investigation and learning. Active participation in the class is very important. This class emphasizes the opportunity for individual discovery rather than mastering a fixed set of techniques. By the end of this class, you are expected to have the skills to answer data analysis questions on target data characteristics of interest, select appropriate regression tools to address these questions, perform the implementation, and interpret the results within the context of the data analysis scope. Regression modeling is the foundation of most other statistical and machine learning modeling. Understanding the regression models introduced in this course will support your analytics skills in other areas. To this day, regression modeling remains the most used data analysis approach. It is worth investing time to understand the basics, as you will use it frequently if you pursue a career in analytics.

What activities will the course involve students in to help them practice and demonstrate their learning?

Homework Assignments: There will be four assignments, including both concept and data example problems. Conclusions and interpretation of results are very important in statistical modeling. These assignments are intended to help you prepare for the midterm and final project. You are allowed (and encouraged) to ask questions about the assignments and collaborate with fellow learners, although you must think through potential solutions on your own and submit your own work. Do not consult any prior homework solutions or resources providing solutions to the homework assignments. It is important to use the homework assignments to practice with the course material and to prepare for the three exams. Please read carefully all our communication on the Honor Code and the use of Generative AI tools.

Midterm Exams: There will be two midterm exams and a final project, featuring problems that review the material (lectures and assignments) provided in this course. The exams are designed to help students grasp standard regression analysis methodology, facilitating a deeper understanding in the application context. The exams are divided into two parts: Part 1 consists of multiple-choice questions, and Part 2 includes data analysis questions. **Both parts will be open-book, following the policy stated below. Please read carefully. Not adhering to this policy will be considered an honor code violation.**

Open-Book Policy on Exams:

What we mean by open-book on the exams is that students are allowed to refer to course materials, including lectures, homework assignments, and any material provided in the course. Open-book includes any notes you may have stored on your computer on course topics; make sure you store the course material on your computers, not online systems (e.g., do not store on Google Colab). **Only one computer is allowed during exams. Open-book in this course does NOT include access to the internet or communication by any means**. **However, the policy will allow the use of Stack Overflow as the only online site you may visit during the exam to get help with coding issues if you get stuck. It does NOT allow the use of Generative AI tools or other external resources. Use of the internet and/or communication with anyone during the exam will be subject to the Georgia Tech honor code and conduct policies/actions** (https://www.policylibrary.gatech.edu/student-life).

Please note that taking the exam in Canvas using the Honorlock proctoring application limits students' use of various platforms. For example, it does not support the use of IDEs such as RTVS, Emacs with ESS, Eclipse with StatET, Sublime Text, Visual Studio, among others. File formats such as QMD are also not supported. The use of virtual machines is not supported. Make sure to set up the exam in an environment that follows the exam guidelines and is supported within the context of the examination system.

Final Project:

The general goal of the final project is to provide you with experience in applying regression analysis methodology to real data. For this project, your team will find a dataset and use that to address the

topics provided for the final project. Your team will write a final report on the analysis addressing these questions. This project will serve as a means for students to demonstrate what they understand and can do with the course material, but it is also recommended to go beyond that. The course material provides fundamentals of regression analysis, however there is much more that you could build on these fundamentals. In grading, we will primarily look for a sensible approach to the problem, and clearly-made connections between your analysis and the substantive questions to be addressed within the project topic. You can use any computing equipment and any computing resources, any written source material you can find, in or out of the school. However, replicating results which have been already published without referencing to the source of publication is subject to plagiarism. Plagiarizing is defined by Webster's as "to steal and pass off (the ideas or words of another) as one's own: use (another's production) without crediting the source." Be sure to document carefully your project work and cite any external materials you may use.

Grading Policy:

In grading the data analysis problems in the exams, we will primarily look for a sensible approach to the problem, and clearly-made connections between your analysis and the substantive questions.

How will students be evaluated?

The course will be letter graded. The grade for the course will be based on two midterms, one final project, and assignments during the semester - Midterm 1: 25%, Midterm 2: 25%, Final: 35%, Assignments: 15%.

The final course grade will be converted into letter grades as follows:

A – 90 to 100

B- to 79 to 89

C – 66 to 78

D – 50 to 65

F – below 50

We will round the scores, that is, a score of 89.50 will become 90 (thus a letter grade A) but 89.49 will become 89 (thus a letter grade B). We also compare the grades from the current semester with those from previous semester to make sure the current semester doesn't have particularly hard assignments and examinations.

Honor Codes and Student Conduct:

All course participants (instructor, teaching assistants, staff and learners) are expected and required to abide by the policies of the Georgia Tech Academic Honor Code, and the Student Conduct expectations (http://www.policylibrary.gatech.edu/student- life). Keep in mind:

- Ethical behavior and personal integrity are extremely important in all facets of life.
- Learners are responsible for completing their own original work. If external resources outside of the course material (including solutions to prior homework assignments and Generative AI tools) or collaboration with other students are to be used in homework assignments, they need to be referenced properly**. <u>Lack of a reference citation is a violation of the honor code.</u>**

- Use of Generative AI tools and any other internet resources will NOT be allowed during the exam, when students will need to demonstrate their learning without relying on external resources. **Use of such resources is a violation of the honor code.**
- Any course participant found in violation of the Georgia Tech Academic Honor Code and/or the Georgia Tech Student Conduct expectations will be subject to the following consequences: 1. Institute a disciplinary warning and assign a grade of zero for the assignment or exam; and 2. Forward the resolution to Georgia Tech's Office of Student Integrity.

Use of Generative AI Tools:

Generative AI tools are now becoming a more integrative part of how we derive knowledge. However, they are a two-edged sword. While they provide opportunities for learning, they also hamper self-learning in a way that new knowledge might not be solidified for understanding new concepts and replicating rigorous analysis. Such tools may also interfere with the development of accurate knowledge since when such tools don't know the answer will make up an answer! Also relying on such tools for learning will particularly impact the employment of your skill set to derive knowledge – it raises the bar of knowledge as highlighted by the Professor Chris Dede. See the link below for more information:
https://www.gse.harvard.edu/ideas/edcast/23/02/educating-world-artificial-intelligence

In this course, we will treat Generative AI tools similar as collaboration with other people: you are welcome to talk about your work with other peer students as well as with AI-based assistants. However, **all work you submit must be your own**. You should never include in your assignment anything that was not written directly by you without proper reference.

Using Generative AI tools could be useful in this course as follows:

- Inquiries about (basic) concepts in the course to complement the explanation of these concepts in the course material. You will have to thread this carefully since the information provided by such tools may not be accurate/correct/rigorous. Do not use these tools as your only approach to complement learning.
- Inquiries about the use of R and python commands that could help in speeding up the process of learning data analysis implementations. The course material provides the coding needed to understand the course material, but such tools may provide additional support that could improve your use of the course material, for example, better ways to develop visual analytics, or use of python commands that could translate the R code provided in the course.

It is unavoidable that such tools will be part of the students' learning hence it is expected that you will consult Generative AI in some instances that could enhance your learning. It is, however, **the students' responsibility** to assume that the information from AI tools will not always be accurate thus you will need to check with your instructor team and/or other resources.

Students need to be aware of the potential harm to their learning as follows:

- Inquiries on developing R or python code for data analysis or on providing interpretations of the data analysis could result in code commands or interpretations that are superfluous, and don't necessarily give a rigorous answer to the problem at hand. Such inquiries will not help students understand what each command is meant to do thus not being able to replicate rigorous data analyses on their own. It is thus important **NOT** to use Generative AI tools to respond to the homework assignment or exam questions. Submitting the assignment and exam questions to such tools is also not compliant with Georgia Tech policies on sharing the course material beyond the classroom learning. *Thus, we will consider an honor code violation if the responses to assignment and exam questions are generated by such tools.*
- Inquiries on any aspect related to students or instructors, and other individual-level

information are considered violations of the Georgia Tech Guidelines on Generative AI for Privacy and Security. Please read carefully the Georgia Tech Guidelines: https://gatech.service-now.com/home?id=kb_article_view&sysparm_article=KB0043472

While we understand that Generative AI tools may be used to complement learning and to enhance coding skills, students should NOT use such tools for developing solutions to homework assignments or exams or the report analysis and write up. Please read about some heuristics and recommendations at: https://www.cc.gatech.edu/news/new-policies-navigate-role-ai-assistants-cs-courses.

Communication:

Course updates will be sent through the **piazza** platform. Please contact your instructor, teaching assistants, and fellow learners via piazza.

- The course will host a class discussion forum using piazza. Feel free to ask questions and respond to other students' questions to the best of your knowledge; this is a learning community, supporting each other.
- While participation in the discussion forum is not mandatory, students are responsible for knowing the content of all pinned Piazza posts. This is where instructor team will be letting students know about important announcements, changes (if any), etc., thus all students are required to read and know the content of all such posts. It is expected that students will check these posts at least once every 24 hours (with exceptions for national and religious holidays, emergencies, etc.).
- Communicate with instructors, teaching assistants and fellow learners using your name as listed in the student roster. E-communication can be less constructive or less thoughtful than in-person communication. When someone does not introduce himself or herself, it is easier to be less respectful. To avoid sensitive situations, we are asking everyone to post in piazza with their name. Posts made by Anonymous profiles will **not** be responded to.
- Please search piazza for your question prior to posting a new one as it may already be answered.
- Instructors and teaching assistants may not be able to address all piazza communications, so we encourage fellow learners to respond to posts. If there is a delay in instructor or teaching assistant response, please be patient and know there may not be a response if we are in a week with heavy volume and/or if the question has already been addressed in a different post.
- Overall, the discussions will be supervised and monitored by teaching assistants under instructor's guidance.

Netiquette:

Netiquette refers to etiquette that is used when communicating on the Internet. Review the Core Rules of Netiquette.

- When you are communicating via email, discussion forums or synchronously (real-time), please use correct spelling, punctuation and grammar consistent with the academic environment and scholarship.
- *In Georgia Tech's MS in Analytics program, we expect all participants (learners, faculty, teaching assistants, staff) to interact respectfully. Learners who do not adhere to these expectations may be removed from the course.*

Course Topics and Schedule:

Please see accompanying documents on Course Outline and Course Schedule.

Course Technology/Software Requirements:

- Internet connection (DSL, LAN, or cable connection desirable)
- R statistical software (free download; see cran.r-project.org)
- Adobe Acrobat PDF reader (free download; see https://get.adobe.com/reader/)
- Jupyter Notebook (free download; see https://jupyter.org/)