Methods of Data Analysis 1

University of Toronto
Department of Statistical Sciences
STA302H1S Winter 2023

Section details:		**Instructor:**	Mohammad Kaviul Khan
LEC5101:	Tuesday 6pm-9pm	**Course email:**	sta302@utoronto.ca
Classroom:	MC102	**Office Hours:**	TBD

COURSE OVERVIEW

Course Description: The course provides a solid introduction to data analysis with a focus on the theory and application of linear regression. Topics to be covered include: initial examination of data, correlation, simple and multiple regression models using least squares, inference for regression parameters for normally distributed errors, confidence and prediction intervals, model diagnostics and remedial measures when the model assumptions are violated, interactions and dummy variables, ANOVA, and model selection and validation. Statistical software will be used throughout and will be required for the completion of various assessments during the term. The development of strong written communication skills will be emphasized.

Learning Outcomes: Learning Outcomes: By the end of this course, all students should have a solid understanding of both the mathematical theory of linear regression analysis and its application in the form of a data analysis tool. Students should be prepared to show their understanding of the above through

1. application of methods through problem-solving questions;
2. description and explanation of concepts relating to the mathematical theory;
3. derivation and proof of topics based on linear regression concepts and theory;
4. recognizing the importance of assumptions and limitations of linear regression models to gauge when linear models are appropriate to use and to be critical of their results.
5. interpreting the results of an analysis involving linear models for technical and non-technical audiences.
6. explaining statistical concepts and theory of linear models to various audiences as would be required in the job market or collaborative environment.
7. outlining the correct use of linear models in a coherent and reproducible analysis plan.
8. applying and extend linear model theory through completion of problem-solving questions

Pre-requisites: Pre-requisites are **strictly enforced by the department, not the instructor.** If you do not have the equivalent pre-requisites, you will be un-enrolled from the course. Students should have a second year statistics course, such as {STA238, STA248, STA255, or STA261}, a computer science such as {CSC108, CSC120, CSC121, or CSC148} and a mathematics course such as {MAT221(70%), MAT223, or MAT240} or equivalent preparation as determined by the department.

COURSE MATERIALS

Course Content: We have a Quercus course page for sections LEC5101 of this course. All lecture slides, any recordings and materials will be posted on this Quercus course page. Further, any important announcements will also be posted in Quercus. Please make sure to check it regularly.

Textbooks: We will be mostly following *A Modern Approach to Regression with R* by Simon J. Sheather (Springer). This book is freely available as an electronic copy through the University of Toronto Library. We will cover Chapters 1-7, with suggested practice problems selected from this book. Datasets and other resources are available from the textbook's website: http://gattonweb.uky.edu/sheather/book/. However, some of the lecture materials will be covered from some other books such as *Linear Regression Analysis* by Douglas C. Montgomery et.al (Wiley) and *Generalized Additive Models: An Introduction with R* by Simon N. Wood (CRC press, 2017). However, these books are optional and you don't need to buy. Rather the focus should be on the lecture materials.

Statistical Software: We will be using the R Statistical Software for performing statistical analyses in this course. R is a free software that can either be downloaded onto your personal computer or used in a cloud environment. We encourage all students to use RStudio through the JupyterHub for University of Toronto. This will allow you to login with your official UofT credentials and use RStudio without the need for a local installation and can be run on any device that has access to an internet connection. More information about using RStudio in JupyterHub will be provided early in the term. R code shown in class will be available on the course page and, along with any additional resources, should be sufficient to complete any assessment involving data analysis.

COURSE COMPONENTS

Lectures: Lectures will be conducted in person in MC102. Slides will be available after the class. Class time each week will comprise of a combination of lecturing, and code-along sessions. Where possible, you are encouraged to bring a laptop or tablet to follow along with the code.

Office Hours: Instructor and TAs will hold office hours in a combination of online and in-person formats. The office hour schedule and mode of delivery will be posted on Quercus once finalized. It is recommended that you visit office hours whenever you have a question about the material. It is always important to have material clarified as quickly as possible. Don't wait until the last minute to ask your questions!

Piazza: We will be using the Piazza as an online discussion forum, which can be accessed through the Quercus course page. **All questions about course material should be posted here** or asked during TA/instructor office hours. The instructor and TAs will monitor the board and will help answer questions but students are encouraged to answer posts and help their fellow classmates.

COMMUNICATION

How your instructor will communicate with you: All communication will be made through Quercus announcements or during lectures. Please ensure that you check Quercus regularly so you don't miss anything important.

Where to send content questions: We will be using the Piazza to collect student questions regarding course content, assignments, etc. All questions should be posted here.

When to email the instructor: The instructor will only respond to emails of a private or sensitive nature. If you email the instructor with content related questions, you will be asked to repost your question on the content board so the answer may benefit all students. Should you need to email the instructor about a sensitive or personal nature, please use your official mail.utoronto.ca email, include your full name and student number in the text. Send all course related emails to sta302@utoronto.ca. Please allow up to 48-96 hours for a reply. Emails will not be monitored on evenings and weekends.

A note on email and discussion board etiquette: Please make sure that you communicate politely

and respectfully with all members of the teaching team and your fellow classmates. Written communications can sometimes take a tone other than what was intended (e.g. can come off as dismissive, rude or insulting), so make sure you re-read or read out loud your email/post before sending it to make sure it has the tone you intended. For more tips on respectful communication, see professional communication tips. Piazza is a teaching and learning tool and therefore should only be used as such. Any posts that detract from the learning goal of the board will be removed to keep the board a safe space.

GRADING SCHEME

All the students will be evaluated in the following way:

Assessment	Date	Weight
Assignment	February 7	10%
Term Test	February 28	25%
Final Project Proposal	March 14	10%
Final Project Report	April 4	25%
Final Exam	April 11-28	30%

Please note that the last day to drop the course without penalty is March 19, 2023.

EVALUATION BREAKDOWN

Assignment: You will be given one assignment in the term. The purpose of this assignment is to develop your understanding of the statistical properties of the estimators obtained from a linear regression model. This will be useful for developing data analysis skills as well as to develop practical understanding of the methods taught in the class. The assignment will have a heavy focus on the use of statistical software (R specifically), and will involve applying the methods learned during lecture to a data set. The format of the assignments will be as follows:

1. use the methods taught in lecture to perform a small data analysis.
2. simulate unique datasets and writing your own functions instead of built in R functions.
3. solve some mathematical problems and explain the procedure

Term Test: The term test will be conducted in person during the scheduled Tuesday class time (see top of page 1). The test will be approximately 2 hours long. More details will be communicated closer to the test date. The test will cover material from Weeks 1-6.

Final Project: The final project will be due on the last day of the lectures and will consist of a data analysis on **a novel dataset** of your choice. Students will be required to demonstrate their understanding of the methods taught in lecture by developing a reasonable regression model that addresses a valid research question using the techniques taught in class. The students will be responsible for choosing the correct methods to apply and providing appropriate justifications defending their choices. The final project is a scaffolded assessment involving 2 parts:

- Part 1- Research question and dataset selection: Students must find a dataset available online and define a research question that can be answered with this dataset using linear regression. Students will need to explain why their research question is important and how linear regression may be used to answer it. A short exploratory data analysis of the chosen dataset will also be required. More details will be provided during the lectures. This part will be due on March 14th.
- Part 2 - Final Project Report: Students will put together a scientific report that outlines the relevance of their proposed research question, the process of their analysis, the results of the performed data analysis, and a discussion of the meaning of the results as well as limitations of the analysis with respect to the statistical tools used/decisions made or the data used. This part will be due on April 4th (the last day of the lectures).

The final project will be done individually, and must be typed and submitted by the deadline. More detailed instructions will be provided at a later date.

Final Exam: The details about the final exam will be provided during the last week lectures. For the final exam we will be following standard University of Toronto Schedule. the final exam will be 3 hours in duration and will be scheduled by the Faculty of Arts and Science during the final assessment period.

LATE ASSESSMENT AND EXTENSION REQUEST POLICY

The assessment deadlines may change from the ones stated in the syllabus depending on how the lecture progresses. However, once the deadline(s) has been announced, the students need to submit the assignments by the deadline. Students will be able to still submit the assignments up to 5 days after the deadline, however, each additional day will be accounted for 20% penalty.
Extreme Situations/Prolonged Illness Extensions: Should a student be experiencing a prolonged illness or other situation that prevents them from turning in their work by the deadline, they should **immediately contact their instructor and College Registrar** to inform them of their situation. They should also submit an Absence Declaration form on ACORN that lists every day during which they were incapacitated and unable to work. Accommodations or further extensions will not be considered without a completed declaration, and will only be considered for extreme circumstances.

Accessibility-Related Extension Requests: Students registered with Accessibility Services should notify the instructor as soon as possible if additional time is needed on assessments that are eligible for extensions. Please **notify the instructor by email of your situation and cc your accessibility advisor** in the process. The instructor will work with the accessibility advisor to determine an appropriate extension for your situation.

MISSED ASSESSMENT POLICY

If you experience a prolonged absence due to illness or emergency that prevents you from completing any number of assessments, please contact your College Registrar as soon as possible so that any necessary arrangements can be made.

Missed Assignment or Final Project: Missing assessments will receive a 0.

Missed Term Test: If a student is experiencing a serious personal illness or emergency on the date of the test, the student **must declare their absence on ACORN and notify the teaching team via email no later than one week after the date of the test**. A make-up test will then be scheduled at a date and time determined by the instructor. **The format of the make-up is at the discretion of**

the instructor and may not resemble the format of the original (e.g. an oral exam).

REGRADE REQUESTS

Regrade requests will be accepted for all assessments. Regrade requests must provide a justification for where there exists a grading error and/or how the work meets the grading rubric. These justifications must further be backed up with concrete references to the course material. All regrade requests will be accepted through a form available on the Quercus course page and will be accepted no later than one week after the grade for that assessment is released. **No regrade requests will be accepted by email or after the 1 week deadline.** The instructor further reserves the right to re-evaluate the assessment in its entirety (i.e. grades can go up, down, or remain unchanged). Please allow a few weeks for regrade requests to be processed by the instructor.

INTELLECTUAL PROPERTY

Course materials provided on Quercus, such as lecture slides, assessments, videos and solutions are the intellectual property of your instructor and are for the use of students currently enrolled in this course only. Synchronous sessions will be recorded and be made available to other students enrolled in the course. **Providing course materials to any person or company outside of the course is unauthorized use and violates copyright**.

ACADEMIC INTEGRITY

The University treats cases of plagiarism and cheating very seriously. It is the students' responsibility for knowing the content of the University of Toronto's Code of Behaviour on Academic Matters. All suspected cases of academic dishonesty will be investigated following procedures outlined in the above document. If you have questions or concerns about what constitutes appropriate academic behaviour or appropriate research and citation methods, you are expected to seek out additional information on academic integrity from your instructor or from other institutional resources (see http://academicintegrity.utoronto.ca/). Here are a few guidelines regarding academic integrity:

- Being dishonest when reporting an illness or personal emergency to get an extension or accommodation is an academic offence.
- You may consult class notes/lecture slides during assessments, however sharing or discussing questions or answers with other students is an academic offence.
- Students must complete all assessments individually. Working together is not allowed unless otherwise specified.
- Paying anyone else to complete your assessments for you is academic misconduct.
- Completing assessments for another student is academic misconduct.
- Sharing your answers/work/code with others is academic misconduct.
- Using sources external to the course (anything not on Quercus) on an assessment is an academic offence.
- All work that you submit must be your own! You must not copy mathematical derivations, computer output and input, or written answers, etc. from anyone or anywhere else. Unacknowledged copying or unauthorised collaboration will lead to severe disciplinary action, beginning with an automatic grade of zero for all involved and escalating from there. Please read the UofT Policy on Cheating and Plagiarism, and don't plagiarise.

ACCESSIBILITY NEEDS

The University of Toronto offers academic accommodations for students with disabilities. If you require accommodations, or have any accessibility concerns about the course, the classroom, or course materials, please contact Accessibility Services as soon as possible: accessibility.services@utoronto.ca or http://accessibility.utoronto.ca.

CIA's University Accreditation Program and Pathway to Actuarial Credential

This course is one of the mandatory courses under Canadian Institute of Actuaries (CIA)'s University Accreditation Program (UAP). UAP has moved away from the course-by-course accreditation method and towards program accreditation method (the "Pathway 1 of CIA qualification"). Under the new pathway, in order to obtain ACIA (Associate of CIA) professional credential, students need to:

1. Complete a degree from an actuarial program (ACT Specialist or Major) at University of Toronto and pass a list of mandatory courses. No minimum course grade or GPA is required as long as students pass all the mandatory courses. The full list of UofT's 16 mandatory courses are: ACT240, ACT245, ACT247, ACT348, ACT349, ACT370, ACT451, ACT452, ACT466, STA257, STA261, STA302, STA314, ECO101, ECO102, MGT201/RSM219. For transition: CIA will accept an actuarial degree from UofT completed between June 30, 2015 and October 31, 2023 without all the specified mandatory courses.

2. Complete the ACIA Module (administered by CIA, projected Spring 2023). For transition: a student can be exempt from the ACIA Module if they complete SOA exam PA and the 8 FAP Modules and assessments by December 31, 2023.

3. Complete an open-book ACIA Capstone Exam (administered by CIA, projected Fall 2023). For transition: a student can be exempt from the capstone exam by completing any combination of UAP credits or exams for P, FM, IFM, LTAM, STAM and SRM by October 31, 2023. **The deadline to apply for UAP credits is September 30, 2023**.

Details on the new pathway for students can be found here: https://education.cia-ica.ca/acia/acia-for-accredited-university-students/.

TENTATIVE SCHEDULE OF TOPICS

Below is a tentative schedule of topics to be covered in class. The schedule is subject to change and modification.

Week (Dates)	Content
1 (Jan. 10)	**Introduction and Good Data Practices:** syllabus overview, review of statistical materials, importance of clear and reproducible communication, subjectivity of statistical tools, good data exploration, good communication practices, intro to JupyterHub and RMarkdown,
2 (Jan. 17)	**Introduction to Modeling:** Inference on the slope and intercept, confidence intervals and prediction intervals, ANOVA , coefficient of determination, indicator variables.
3 (Jan. 24)	**Multiple linear regression:** motivation through polynomial regression, review of matrix linear algebra, parameter estimation in MLR, properties of least squares estimates.
4 (Jan. 31)	**Diagnostics for Simple Linear Regression:** residuals and residual plots, leverage and influential points .
5 (Feb. 7)	**ANOVA and ANCOVA:** Confidence intervals for parameters, F-test, partial F-test, working with indicator/dummy variables . **Assignment due**
6 (Feb. 14)	**Handling violations in Simple Linear Regression:** transformations to stabilize variance, transformations for non-linearity, Box-Cox.
Feb. 21-24	READING WEEK
7 (Feb. 28)	**Term Test**
8 (Mar. 7)	**Diagnostics for Multiple Linear Regression:** residuals and their properties, standardized residuals, leverage points, residual plots, influential observations.
9 (Mar. 14)	Multicollinearity. Weighted and Generalized Least Squares in Multiple Linear Regression Handling violations and Variable Selection. **Project Proposal due**
10 (Mar. 21)	**Variable selection:** variable selection procedures, model validation, Ridge regression and LASSO
11 (Mar. 28)	**Data Reduction:** Principal Components Analysis. Orthogonal Variables
12 (Apr. 4)	**Beyond Linear Regression:** Generalized Additive Models and Generalized Linear Models. **Project Report due**
Apr 11-28	Final assessment period