Methods of Data Analysis 1

University of Toronto
Department of Statistical Sciences
STA302H1F Summer 2024

Section details:		**Instructor:**	Dr. Antonio Herrera Martin
LEC Mon, Wed:	2-5pm	**Course email:**	sta302@utoronto.ca
Classroom:	B150 in Pharmacy Building (BP)	**Office Hours:**	Tuesdays 15:00-16:00 (in person)

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, ANOVA, and model selection and validation. Statistical software will be used throughout. The development of strong written communication skills will be emphasized.

Learning Outcomes: By the end of this course, all students should be able to:

1. Recognize 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.
2. Interpret the results of an analysis involving linear models for technical and non-technical audiences.
3. Apply methods of linear models and data cleaning to new datasets correctly using statistical software in a reproducible way.
4. Explain statistical concepts and theory of linear models to various audiences.
5. Outline the correct use of linear models in a coherent and reproducible analysis plan.

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 this course. Handouts 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.

Textbook: This course does not strictly follow any particular textbook, but rather merges material from a number of sources. **All of the below recommended textbooks are freely available as an electronic copy through the University of Toronto Library.** Our primary reference text will be

- *A Modern Approach to Regression with R*, by Simon J. Sheather (Springer)

Other helpful references from which practice problems may be assigned are:

- *Applied Regression Modeling*, 2nd edition, by Iain Pardoe (Wiley).
- *Methods and Applications of Linear Models*, 2nd edition, by Ronald R. Hocking (Wiley)

- *Applied Linear Regression*, 3rd edition, by Sanford Weisberg (Wiley).

These are all useful books, but may present the material in a different order or in a different way. They are still good for additional explanation and practice problems. Other useful resources will be posted on the Quercus course page.

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 B150 in the Leslie L. Dan Pharmacy Building (BP). It will cover material from the handouts, and if appropriate slides will be available after by the end of the week's class. Class time each week will comprise of a combination of lecturing, in-class activities, 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!

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.

Piazza: Piazza will be used as an online discussion forum, which can be accessed through the Quercus course page. The discussion board will be peer-monitored, and students are encouraged to answer posts and help their fellow classmates answering questions about course material. If there are persistent question, they should be asked during TA/instructor office hours. The TAs will not be monitoring the board, but the instructor will sporadically monitor, so that university and course etiquette is respected.

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, student number and lecture section (e.g. L0101) in the text. Send all course related emails to sta302@utoronto.ca. Please allow up to 48 hours for a reply. Emails will not be monitored on evenings and weekends.

A note on email and discussion 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. 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

Assessment	Date Due/Occurring	Scheme
Assignment 1	May 15	10%
Assignment 2	May 22	10%
MidTerm Test	May 27	20%
Final project	June 12-17 TBD	30%
Final Exam	June 19 – 24	30%

Please note that the last day to drop the course without penalty is June 3, 2024.

EVALUATION BREAKDOWN

Assignments: You will be given two assignment in the term. The purpose of these assignments is to describe your understanding of a linear regression model. 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. During the first assignment students will solve a mathematical problem and explain the procedure by hand.
The second assignment will be an automatic quiz which will be used to asses the understanding of results.

Term Test: The term test will be conducted in person during the scheduled class time for all sections. The test will be 2 hours long. More details will be communicated closer to the test date. The test will cover material from Weeks 1-3.

Final Project: The final project will be due during the last weeks of term (date to be confirmed as soon as possible) 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 to defend their choices.

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.

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.

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.

The final project 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 2-3 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

Assessments and project: Students will be able to submit the first assignment and final project up to 3 days after the deadline, however, each additional day will be accounted for 10% penalty. The second assignment will be in the form of a online quiz and will close at the due date, it will allow 3 attempts before the due date, so no late assessment will be allowed.

Extreme Situations/Prolonged Illness Extensions: Should a student be experiencing a prolonged illness or other situation that prevents them from turning in their work, 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 MidTerm 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**. The proportion of the mark of the midterm will be transfered to the final exam which will be **50%.**

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 will be accepted by email and will be accepted no later than one week after the grade for that assessment is released. No regrade requests will be accepted 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.

USE OF ARTIFICIAL INTELLIGENCE

The latest generation of Artificial Intelligence is impacting teaching and learning, while it presents important opportunities for learning, it also needs to be taking in consideration the appropriate use within the course for fair evaluation and learning. For this course, these are the main guidelines of the use of AI tools, unless otherwise specified in the instruction of an assignment:

- **Students may use artificial intelligence tools for critiquing and editing an assignment for purposes of revision, but the first draft must be original work produced by the individual student alone.**
- **Students may use artificial intelligence tools for creating an outline for an assignment, but the final submitted assignment must be original work produced by the individual student alone.**
- **It should be included what tool(s) were used, how they were used, and how the results from the AI were incorporated into the submitted work.**
- **Students may not use artificial intelligence tools for taking tests in this course.**
- **Course instructors reserve the right to ask students to explain their process for creating their assignment.**
- **If you have any question about the use of AI applications for course work, please speak with the instructor.**

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.

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	Content
1	Introduction syllabus overview, Matrice and Vector, statistics overview, intro to JupyterHub and RMarkdown/Quarto, Modelling, Simple linear regression, Model coefficients, Errors and Maximum, WLE
2	Linear Regression(cont), Confidence and test in regression Multiple linear regression, Sampling distributions, Confidence intervals, prediction intervals estimates of error variance.
3	Decomposing the Variance and mitigating violated assumptions Sum Squares, ANOVA, ANCOVA, F-tests, Multicolinearity, Plot diagnostics, Transformations and non-linearity

4	**Midterm and Diagnostics: Midterm test and Model adequacy, Outlying observation**
5	**Regression models for quantitative and qualitative predictors Polynomial regression, Interactions, Qualitative predictors; Ridge Regression.**
6	**Autocorrelation in time series: Autocorrelation and AR(1)**