Methods of Data Analysis 1

University of Toronto
Department of Statisitical Sciences
STA302H1S Summer 2024

Section details:	**Instructor:** Yaoming Zhen, Ph.D.
LEC 5101 Mon. & Wed. 6pm-9pm	**Course email:** sta302@utoronto.ca
Venue: MS 3153	**Office hours:** Tue. 4pm-5:30pm

1 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: By the end of the 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 as would be required in the job market or collaborative environment.

5. Outline the correct use of linear models in a coherent and reproducible analysis plan.

6. Distinguish the common and unique characteristics between linear regression model and any other given machine learning models, which is important for learning or designing a new model.

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 (STA238H1/ STA248H1/ STA255H1/ STA261H1/ ECO227Y1/ STAB57H3/ STA258H5/ STA260H5/ ECO227Y5), a computer science course (CSC108H1/ CSC110Y1/ CSC120H1/ CSC148H1/ CSCA08H3/ CSCA20H3/ CSCA48H3/ CSC108H5/ CSC148H5) and a mathematics course (

MAT223H1/ MAT224H1/ MAT240H1/ MATA22H3/ MATA23H3/ MAT223H5/ MAT240H5/ MATB24H3/ MAT224H5), or equivalent preparation as determined by the department.

2 Course materials

Course Content: We have a Quercus course page for this course. All lecture slides, assignments and any other materials will be posted on this Quercus course page. In addition, 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 materials 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: Regression and the Analysis of Variance, 2nd edition, by Ronald R. Hocking (Wiley), and

- 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.

3 Course components

In-person class: In-person classes occur on Monday and Wednesday evenings (see ACORN for room), and they mainly consists of lectures and tutorials. In the first two and a half hours, the instructor will

deliver a lecture covering the materials in the slides, and teaching assistants (TAs) will provide tutorials on application and implementation related to the contents of the lecture in the remaining half an hour. Breaks will be taken appropriately during the evening. Where possible, you are encouraged to bring a laptop or tablet (or any device that can connect to a web browser) for hands-on practice.

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!

Quercus discussion board: We will use the Quercus discussion board 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 instructor or TAs 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.

4 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.

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. Include your lecture section (L5101) in the subject line so it is received by the correct person. 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 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. The Quercus discussion board 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.

5 Grading scheme

Each student's final grade will be computed according to the below grading scheme. No special rounding rules or individual grade adjustments (e.g. to meet GPA cut-offs, minimal requirements for programs, etc.)

will be used to calculate course grades. No special reweighting of assessments or extra work will be accepted to account for perceived poor performance, nor to account for any assessment(s) that have been missed without accommodation. There are no exceptions to these policies.

Assessment	**Date Due/ Occurring**	**Weight**
In-class quizzes		
Quiz 1	Lecture 2 (July 8)	2%
Quiz 2	Lecture 4 (July 15)	1 %
Quiz 3	Lecture 9 (July 31)	2%
Midterm		
Midterm Test (during scheduled class)	July 22	25%
Project		
Research proposal	July 17	5%
Final report	August 14	20 %
Final exam (during final exam period)	Scheduled by the FAS	45%

Please note that the last day to drop the course without penalty is July 29, 2024.

6 Evaluation breakdown

In-class quizzes: Students will have 3 in-class quizzes via Quercus as scheduled above. Therefore, make sure to bring your electronic device that can assess Quercus. At every quiz, students will be given 10-15 minutes to finish 2 to 3 problems. The problem formats will be individual or multiple choices, true or false, blank filling, matching and others. It usually does not contain heavy computations or derivations, but it is more about the methodologies and conceptual understandings. You will get full marks as long as you answer 1 out of 2, or 2 out of 3 questions correctly.

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 Lectures 1 - 5.

Final Project: The final project will consist of a data analysis on a dataset. Students will be required to demonstrate their understanding of the methods taught in the course by developing a reasonable regression model that addresses a valid research question using the techniques from the course. 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 two parts:

Research proposal: Students will be tasked with defining a research question that can be answered with a dataset using linear regression. This portion of the project will require students to provide their research question, explain why linear regression would be a reasonable method to answer this question, and highlight important characteristics of their dataset.

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.
All parts of the final project must be done **in groups of 4 to 5 students**. All group members are expected to contribute to the project equally and provide an outline of their involvement in the project. More detailed instructions for each part will be provided on Quercus 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.

7 Late assessment and extension request policy

Extensions of the final project: All groups for the final project will have access to extensions of up to 7 days to help manage illness, deadlines or other unexpected situations, but they need to inform the instructor with justification. Groups may use these extensions on the research proposal or final project report, but they need to inform the instructor in advanced or at most 7 days after the required submission deadline. While using these extensions for your final project, all group members must agree to use this extension, so groups should strive to have clear communication throughout the term. Groups who turn in the work by the assigned deadline (i.e. do not use the extension) will receive their graded work and feedback earlier than groups who use extensions. Extensions beyond this will not be granted.

Extreme Situations/Prolonged Illness: Should a student be experiencing a prolonged illness or other situation that prevents them from turning in their work or contributing to the group project, 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 or a Verification of Illness (VOI) form that lists every day during which they were incapacitated and unable to work. These documentations should be sent to sta302@utoronto.ca. Accommodations will not be considered without a completed declaration or VOI, and will only be considered for extreme circumstances at the request of the College Registrar.

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 such accommodation. 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 accommodation for your situation. However, note that group work can generally not be granted further extensions beyond those in the above policy.

8 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 In-classes quizzes: There will be no accommodations made for missing the in-classes quizzes.

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.** If a student misses the term test for a valid reason then the weight of the term test (25%) will be shifted to the weight of the final exam. In such case, the weight of the final exam will be 70%.

Missed Final Project: Due to the nature of these assessment, there will be no further extensions on the research proposal or project (see Extensions on the final project in Section 7) under any circumstances. Late projects will not be accepted and there are no accommodations available for individuals missed contributions to their group projects.

9 Regraded 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.

10 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. In class lectures and tutorials might 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 unauthorizeduse and violates copyright.**

11 Use of artificial intelligence

ChatGPT and other generative AI are freely available tools that can perform a variety of functions for us. However, it's important to understand how such tools are allowed to be used in this course. Acceptable uses of generative AI in this course include:

- Editing or rephrasing written work that has already been written by the student to improve the syntax, grammar and overall readability of the work.

- Synthesizing or explaining course concepts while learning and studying to contribute to their understanding of the course material.
- Looking up appropriate syntax of individual R functions for use in a data analysis or for understanding errors that may arise when running R code.

However, the work turned in by students must ultimately be their own and students will therefore be accountable for the work they turn in. Unacceptable uses of generative AI in this course include:

- Copying from any generative artificial intelligence applications, including ChatGPT and other AI writing and coding assistants, for the purpose of completing assignments in this course.
- Producing an entire data analysis, written report, or any other piece of work meant for grades.

In summary, generative AI like ChatGPT can be really helpful in your learning process and to improve skills valued in the workplace. However, it cannot be used as a substitute for learning and material produced from these tools should not be passed off as your own. This would be considered academic misconduct (see below). The instructor therefore reserves the right to ask students to explain their work and their process for creating their assignment.

12 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 take-home 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.

13 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 coursematerials, please contact Accessibility Services as soon as possible: accessibility.services@utoronto.ca or `http://accessibility.utoronto.ca`.

14 Tentative schedule of topics

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

Lecture (Dates)	Content
1 (July 3)	**Introduction** of the course, data formats, machine learning methods, and **Simple linear regression basics:** functional and statistical relationship, simple linear regression model, least square estimation, and interpretation.
2 (July 8)	**Multiple linear regression basics:** vector and matrix operations, multiple linear regression model, least square estimation, and interpretation.
3 (July 10)	**Assumptions of linear regression:** assumptions, residual plots and violation detection.
4 (July 15)	**Correcting assumptions:** Transformations, sampling distributions, and the first slight of polynomial regression and generalized linear models.
5 (July 17)	**Inference in linear regression:** hypophesis tests, confidence intervals, and prediction intervals.
6 (July 22)	Midterm test, multi-collinearity, and a very brief touch on penalized regression.
7 (July 24)	**Decomposition of variance I:** sum of squares decomposition, ANOVA, and F test.
8 (July 29)	**Decomposition of variance II:** partial F-test, coefficients of determination.
9 (July 31)	**Problematic observations:** outliers, leverage points, influential points, detection and influences.
Aug. 5	Civic holiday - University closed; no classes.
10 (Aug. 7)	**Model building and variable selection:** interaction, best subset selection, forward and backward variable selection
11 (Aug. 12)	**Model validation**: validation, cross-validation, report writing workshop
12 (Aug. 13)	Make-up lecture if necessary, review or Q& A
Aug. 15-23	Final assessment period