# Methods of Data Analysis 1

University of Toronto
Department of Statistical Sciences
STA302H1F Fall 2024

| | | | |
|---|---|---|---|
| **Section details:** | | | |
| LEC0101/2001: | Fridays 11am-1pm | **Instructor:** | Dr. Katherine Daignault |
| LEC0201/2002: | Fridays 3pm-5pm | **Course email:** | sta302@course.utoronto.ca |

## 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 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 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. Apply 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 common Quercus course page for sections L0101/2001 and L0201/2002 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.

**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 two primary reference texts will be

- *Linear Models in Statistics*, 2nd edition by Alvin C. Rencher and G. Bruce Schaalje (Wiley).
- *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 will be made available on the course page and, along with any additional resources, should be sufficient to complete any assessment involving data analysis.

## COURSE COMPONENTS

**Asynchronous Online Pre-Class Material:** Prior to each Friday lecture, students will be responsible for working through a Guided Practice Module that includes watching short lecture videos (less than one hour), working on practice problems and submitting these to a Quercus Quiz. This pre-class work covers the basic course content and is necessary for the in-person class activities during class. Each module will be available on Friday evening following lecture, usually no later than 8PM ET.

**Synchronous In-Person Class:** In person classes occur on Fridays (see ACORN for room) following completion of the online pre-work. These classes will consist of a combination of short lecturettes, activities and labs. Where possible, you are encouraged to bring a laptop or tablet (or any device that can connect to a web browser) to class for labs, however most activities can be done in pairs. The in-person classes will solidify the knowledge from the pre-work through practice and application with the support of the instructor and teaching assistants.

**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 be using 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 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 Quercus Discussion board 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. **Include your lecture section (e.g. L0101) in the subject line so it is received by the correct person.** Send all course related emails to sta302@course.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.

## 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** |
|---|---|---|
| Engagement Activities | | |
| Completion of pre-work quizzes (7 out of 10) | Thursdays by 8PM ET | 2% |
| Completion of in-class worksheets (7 out of 10) | Fridays by 8PM ET | 3% |
| Pre-requisite and Syllabus Quiz | September 20 by 8PM ET | 1% |
| In-class Participation | Fridays in lecture | 2% |
| Term Test (all sections) | Oct. 11 from 5PM-7PM ET | 20% |
| Final Project (3 parts) | | |
| Part 1: Research Question/Proposal | October 4 by 8PM ET | 6% |
| Part 2: Analysis Flowchart | October 25 by 8PM ET | 6% |
| Part 3: Written Final Report | November 29 by 8PM ET | 15% |
| Final Exam (during final exam period) | Scheduled by FAS | 45% |

Students who choose not to use the PollEverywhere software will have the In-Class Participation reweighted to the In-Class Worksheets. Both grading schemes will be computed for all students and the higher grade will become the final grade.

**Please note that the last day to drop the course without penalty is November 4, 2024.**

## EVALUATION BREAKDOWN

### Engagement Activities

**Pre-Work Quizzes:** At the end of each guided practice, there will be a set of exercises students should complete either during or after watching the videos. These are meant to provide practice of the basic concepts presented in the videos. Students will be asked to submit their answers to these questions through a Quercus quiz due Thursdays at 8PM ET. Quizzes will only be counted for completion, so students must only answer each question to receive credit. Only 7 out of 10 such quizzes will be counted towards this portion of the grading scheme. The responses to these quizzes will inform whether additional review of the basic concepts is needed in class on Friday. No extensions will be granted, and no accommodations will be made for missed quizzes as only 7 out of 10 will be counted.

**In-Class Worksheets:** Following in-class review of pre-class content, students will work through a worksheet that provides practice in applying the content using R as well as solidifying the concepts using data. Worksheets are to be submitted to MarkUs during or immediately after lecture, or no later than 8PM ET on Fridays. The code is auto-graded and students receive completion credit as long as more than 75% of tests pass. Only 7 out of 10 worksheets will be counted towards this portion of the grading scheme. As such, no extensions will be granted, and no accommodations will be made for missed submissions.

**Syllabus and Pre-requisite Quiz:** There will be 1 short multiple choice quiz early in the term to ensure that students are prepared for the course in terms of their knowledge of prerequisite material and the syllabus content. This quiz will be conducted on Quercus and will be open for students to take at any time until the deadline. Students will get 2 attempts and the highest score will be counted towards their final grade. On each attempt, students will be given 1 hour to complete the quiz, and each question will show up one at a time and will be locked once the question has been answered.

**In-Class Participation:** PollEverywhere will be used to gauge student comfort with the course content and will prompt discussion and further elaboration on the course topics. Students must register using their UofT email address at PollEv.com/katherinedai702 and be signed in to have the participation recorded. Correct answers are NOT required to obtain full credit. The 2% participation grade will be computed based on the percentage of polls a student responds to, following the below scheme. Due to the flexibility, no extensions or accommodations will be made for missed participation.

| | % polls answered | | | | |
|---|---|---|---|---|---|
| | 0 | $(0, 25)$ | $[25, 50)]$ | $[50.75)$ | $[75, 100]$ |
| Grade (out of 2) | 0 | 0.5 | 1.0 | 1.5 | 2.0 |

### Term Test

The term test will be conducted in person on **Friday October 11, 2024 from 5-7PM ET**. The test will be approximately 1 hour and 40 minutes long. More details regarding format will be communicated closer to the test date. The test will cover all material from Modules 1 to 5.

***Any student with an academic conflict (e.g. tutorial, lecture for another class) must contact the teaching team by email NO LATER THAN September 27, 2024 to advise us of this conflict. No accommodations can be guaranteed for any conflict reported after this date.***

**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 three parts:

- Part 1- Research question: 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.
- Part 2 - Analysis Plan Flowchart: Students will be asked to put together a flowchart outlining the steps that they plan to take in their data analysis for the final project on their chosen dataset. This will help in developing a consistent analysis flow and in structuring the methods section of their final report.
- Part 3 - 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 two to three 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.

## LATE ASSESSMENT AND EXTENSION REQUEST POLICY

**'No Questions Asked' Extensions:** All groups for each part of the final project will have access to 'No Questions Asked' (NQA) extensions of up to 7 days to help manage illness, deadlines or other unexpected situations. Groups may use these extensions on any part of the final project **without having to request an extension from the instructor**. The NQA extensions work as follows:

- Students **should not** notify the instructor when using these extensions - we will simply accept the work up to 7 days after the assigned deadline without penalty.
- 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 the NQA extension.
- Groups planning to use these extensions should **only** submit material to Quercus when they are ready to make their final graded submission.
- Extensions beyond this will not be granted.

**Late Submission Policy:** Due to the flexible grading schemes and extension policy, no late submissions will be accepted throughout this course.

**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 BOTH their instructor and College Registrar** to inform them of their situation. The teaching team

cannot support you and your group if we are not made aware of these situations as soon as they occur.

**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.

## 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 Engagement Activities:** There will be no accommodations made for missing the engagement activities. The Pre-work quizzes and in-class worksheets already provide accommodation for missing up to 3 weeks of class. The syllabus quiz will be available for a substantial amount of time and so no accommodation will be offered for this. Students who miss or choose not to complete the in-class participation will be graded according to the secondary scheme.

**Missed Term Test:** If a student is experiencing a serious personal illness or emergency on the date of the test, the student **must complete an Absence Declaration form on ACORN and notify the teaching team at sta302@course.utoronto.ca no later than one week after the date of the test**. Only one make-up test will be available for students who provide legitimate documentation for missing the original test.

IMPORTANT: students may only use the Absence Declaration once per academic term (e.g., the Fall term) for a maximum period of 7 consecutive calendar days. See A&S Student Absences for additional information on eligibility. The once-per-term limit is set by the ACORN Absence Declaration Tool. Once a declaration is submitted, students will be restricted from using the tool to declare any further absences in that term. Students may then submit a Verification of Illness form completed by a physician or a letter from a College Registrar.

**Missed Final Project:** Due to the nature of this assessment, there will be no extensions on any portion of the project under any circumstances. Late projects will not be accepted and there are no accommodations available for individuals missed contributions to their group's project

## 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 ma-**

**terials to any person or company outside of the course is unauthorized use and violates copyright.**

## ACCEPTABLE USES OF GENERATIVE AI

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.

## 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:

- Using ChatGPT and other generative AI for any purpose not outlined above.
- 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.

**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**

UAP has moved away from the course-by-course accreditation and is now based on a program accreditation method. Under the new credentialing pathway, 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. The full list of UofTs 16 mandatory courses are: ACT240, ACT245, ACT247, ACT348, ACT349, ACT370, ACT451, ACT452, ACT466, STA237/STA257, STA238/STA261, STA302, STA314, ECO101, ECO102, MGT201/RSM219;
2. Complete the ACIA online Modules (offered through CIA directly);
3. Complete an online, open-book ACIA Capstone Exam (offered through CIA directly)

For further information on ACIA modules and Capstone Exam, please check CIA's website (cia-ica.ca) or email education@cia-ica.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 (Dates) | Content |
|---|---|
| 1 (Sept. 2-6) | **Simple Linear Regression Basics:** relationships, notation, estimation, interpretation. |
| 2 (Sept. 9-13) | **Multiple Linear Regression Basics** relationship, notation, estimation, interpretation. |
| 3 (Sept. 16-20) | **Assumptions of Linear Regression:** introduction to assumptions, residuals and residual plots, detecting violations, ethics workshop 1. |
| 4 (Sept. 23-27) | **Correcting Assumptions:** Transformations, interpretation, role in sampling distributions. |
| 5 (Sept. 30 - Oct. 4) | **Inference in Linear Regression:** Hypothesis tests and/or confidence intervals on coefficients and mean responses, prediction intervals |
| 6 (Oct. 7-11) | TERM TEST |
| 7 (Oct. 14-18) | **Decomposition Of Variance Part 1:** Sum of squares decomposition, ANOVA F test, Partial F test |
| 8 (Oct. 21-25) | **Decomposition of Variance Part 2:** coefficients of determination, multicollinearity |
| Oct. 28 - Nov. 1 | READING BREAK |
| 9 (Nov. 4-8) | **Problematic Observations:** Outliers, Leverage Points, Influential Points, Detection and Impact, writing workshop 1 (to be confirmed) |
| 10 (Nov. 11-15) | **Model Building and Variable Selection:** Context, likelihood criteria, automated methods, hypothesis tests, ethics workshop 2. |
| 11 (Nov. 18-22) | **Model Validation and Wrap-up:** How to validate your models, MLR data analysis process overview, writing workshop 2 (to be confirmed) |
| 12 (Nov. 25-29) | **Review/Office Hour class:** Final exam review and/or Q&A. |
| Dec 6-23 | Final assessment period |