Methods of Data Analysis 1

University of Toronto
Department of Statistical Sciences
STA302H1S/1001H1S Winter 2021

Instructor:	Katherine Daignault	**Synchronous Class:**	
Course email:	sta302@utoronto.ca	LEC 0101:	T 10AM-12PM EST
Office Hours:	R 11AM-12PM, 4-5PM EST	LEC 0201:	R 1-3PM EST
Course webpage:	Quercus		

COURSE OVERVIEW

How will this course operate? This course will be offered entirely online, with a combination of synchronous lectures and asynchronous video lectures. The majority of course content will be uploaded to Quercus as pre-recorded videos to be watched prior to the synchronous meetings. The synchronous classes will occur through Bb Collaborate in Quercus and will focus on additional materials, worked examples and demonstration of concepts and applications using statistical software. It is your responsibility to make sure you are available during scheduled lecture times and stay on top of the course material and all relevant deadlines.

****** Please ensure that you have access to reliable internet service, as evaluations will take place online and there is no guarantee that accommodations can be provided for faulty internet during an evaluation.***

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, model selection, and penalized regression. Statistical software will be used for illustration purposes and will be required for the completion of various assessments throughout the term.

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. Students should be prepared to show their understanding of the above through

- application of methods through problem-solving questions;
- description and explanation of concepts relating to the mathematical theory;
- derivation and proof of topics based on linear regression concepts and theory;
- practical application of methods on real data using statistical software, with appropriate justification of use of these methods;
- interpretation of data analysis results in clear and non-technical language

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 all sections of this course. All lecture slides, 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: The course closely follows *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 many other freely available textbooks are also recommended, depending on your learning style. Some alternatives from which practice problems may be drawn are:

- *Applied Linear Regression*, 3rd edition, by Sanford Weisberg (Wiley).
- *Applied Regression Modeling*, 2nd edition, by Iain Pardoe (Wiley).

These are both nice books, but present the material in a different order. Good for additional explanation and practice problems.

Statistical Software: We will be using RStudio for performing statistical analyses. R is a free software that can either be downloaded onto your personal computer or used in the cloud. If you choose to work with R on your personal computer, then installation will be a two step process:

1. The base R framework is available for download at `http://cran.r-project.org/` for Windows, Mac and Linux operating systems.
2. Next, RStudio is a good integrated development environment to R (makes it simpler to work in R) and can also be downloaded for free at `https://www.rstudio.com/products/rstudio/download/`.

If you don't want to download the program or run into problems with installation, you may want to consider using 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. More information about using RStudio in JupyterHub will be provided in the first class. All R code and resources for assignments will be provided in lectures and on the course page.

COURSE COMPONENTS

Lectures: The majority of the core content for this course will be delivered via pre-recorded video lectures which will be posted to Quercus each week (ideally by Saturday the latest). It is the student's responsibility to watch these videos in a timely fashion. Synchronous classes will occur through Bb Collaborate in Quercus and will supplement these videos with additional worked examples and data analysis demonstrations. Synchronous lectures will also be recorded and posted a few days after the class.

Office Hours: Instructor and TAs will hold office hours through Bb Collaborate in the Quercus course page. The office hour schedule 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 more important than ever in an online class 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. **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. The board will be organized by week to help keep it all organized.

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 post your question on Quercus so the answer may benefit all students. Should you need to email the instructor, 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.

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.

GRADING SCHEME

Both undergraduate and graduate students will be offered two grading schemes that will be used to calculate your final grade. Your final grade for the course will automatically be determined by the **higher** of the two grading schemes. Both undergraduate and graduate students will have the same grading scheme below:

Assessment	**Date Due/Occurring**	**Scheme 1**	**Scheme 2**
Discussion Board Participation	Sundays every 2 weeks	10%	5%
Assignments (× 3)	Sunday Feb. 14 by 23:59EST	10%	10%
	Sunday March 28 by 23:59EST	10%	10%
	Sunday April 9 by 23:59EST	10%	10%
Mini Project Part 1	Sunday Feb. 7 by 23:59EST	5%	5%
Mini Project Part 2	Sunday March 7 by 23:59EST	15%	20%
Term Test	March 9 and 11 (in class time)	20%	20%
Final Project	Due Sunday April 18 by 23:59EST	20%	20%

MINIMUM PASSING REQUIREMENT

In order for the instructor to be able to reasonably assess the ability of each student with the course material, a minimum amount of work must be submitted to provide enough evidence of proficiency. To this end, students must submit the following assessments in order to be considered for passing the course: **the mini project, the term test, and the final project**. If a student fails to submit the minimum passing requirement, they may not be eligible to pass the course as not enough work has been completed for a meaningful grade to be awarded.

EVALUATION BREAKDOWN

Discussion Board Participation: Participation is mandatory for the graded discussions, but optional for course-content questions, and will be conducted through the use of the Quercus discussion board. Only the graded discussion board counts towards participation marks. The discussion board will be used in two different ways:

- **Ungraded discussion:** there will be a dedicated discussion board where students can post questions regarding course content. The instructor and TAs will monitor this and answer questions posted by students. But it is encouraged that students try to answer students posted from other students. Participation on this discussion board is not mandatory and does not count for grades.

- **Graded participation discussion:** Every two weeks, a discussion topic will be posted based on content presented in the last few weeks of lectures (**see schedule at end of document for exact deadlines**). These topics will require students to discuss various applications of the course material and to think about how and why certain methods may be appropriate or not. All students are encouraged to participate in these discussions for their participation grade. Topics will be open-ended (there is no one right answer) and TAs and instructors will also be involved in these discussions. These will begin the week of January 25 and participation is mandatory. Topics will remain open for contribution for two weeks so it's best not to wait until the last minute to contribute. A rubric will be posted explaining how this will be graded.

Assignments: There will be 3 assignments that must be completed **individually**. These are opportunities to practice and receive feedback on theoretical and applied problems, as well as working with statistical software and demonstrating core course concepts through simulation.

- Since assignments are to be done individually, it is not appropriate to post questions about how to do or approach assignment questions on the discussion board. However, you may ask for clarification, or general content or R questions.

- Assignments will be submitted through Crowdmark, meaning you will need to upload PDF, PNG or JPEG versions of your assignment answers. Instructions will be provided on the Quercus assignment page when posted.

- Punctuality is key to keep the course moving for everyone. There will be a 20% penalty for each day that the assignment is late (e.g. if you submit 15 minutes, 2 hours of 19 hours late, these all receive a 20% late penalty). No assignments will be accepted 48 hours past the due date.

- You will be given 1.5-2 weeks to complete each assignment which should be plenty of time to complete the assessment. In general, no extensions will be granted unless under extreme and unusual circumstances that will be assessed on a case by case basis. If you think you fall into this extreme circumstance category, then you must email the instructor no later than 24 hours BEFORE the due date.

Term Test: The term test will be conducted online. The test will be 1 hour and 30 minutes long, with an additional 20 minutes available to upload your solutions to Crowdmark. A link with the test questions will be emailed to students at the start time of the test, and all submissions must be received before the deadline to be graded. More details on submission will be communicated closer to the test date. The term test will take place during the scheduled synchronous lecture times on **Tuesday March 9 from 10AM-12PM EST and Thursday March 11 from 1-3PM EST**. As per the timetable delivery instructions, students must be available during this time. You will be required to write the term test in the section in which you are enrolled. The test will cover material from Week 1-6.

Mini Project: You will be given a two-part mini project this term. The purpose of the mini project is to develop your data analysis skills which will be useful for the final project and future courses, in addition to your communication and statistical presentation skills. The mini project will have a heavy focus on the use of statistical software (R specifically), and will involve applying the methods learned during lecture to a dataset. The format of the project will be as follows:

- PART 1 will feature a short exercise on communicating complex statistical concepts using common and non-technical language. More details will be provided with the assignment document.
- PART 2 will require you to use the methods taught in lecture to perform a small data analysis and then present your results to a general audience.
- To submit your results for PART 2, you will be required to prepare a 5 minute presentation that you will need to record (using your computer, phone, etc.). You will be required to display your T-card alongside your face at the beginning of your video to verify your identity.
- You will need to display the results of your project in a logical way using slides (e.g. PowerPoint, or other) and record yourself discussing these results, with a focus on why you chose to do certain things and interpretation of your results for a general (non-statistical) audience.
- Presentations should be submitted on time (i.e. by the deadline). Late submissions will receive a 20% penalty for each day that the project is late, up to 48 hours at which point the project will not be accepted.
- In general, extensions will not be given unless under extreme and unusual circumstances. If you think you qualify, your request must be received 24 hours BEFORE the deadline.
- **There is no make-up mini project**. A missed mini project will be given a grade of 0.

Final Project: The final project will be due during the final assessment period **on Sunday April 18 by 23:59EST** and will consist of a data analysis on a novel dataset. Students will be required to demonstrate their understanding of the methods taught in lecture by developing a reasonable regression model that addresses the 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 will be submitted as a project report, which consists of:

- Introduction section: provides details regarding why the model is being developed, general information regarding how the model is developed and finally how the model meets the purpose mentioned earlier
- Methods section: a detailed discussion of the process used to come to the final model, as well as in-depth diagnostics to illustrate the 'goodness' of the model.
- Results section: a presentation of the results outlined in the methods. This section will also feature a comprehensive description of data characteristics using appropriate summary measures, tables and figures that highlight elements of the data that are relevant to building the model.
- Conclusion section: restate why the model is useful in the context of the data, provide an interpretation of the final model in non-technical language, and discuss any limitations/problems remaining with the model and how they might impact its use in the real world.

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. **In order to pass the course, you must submit the final project.**

MISSED ASSESSMENT POLICY

Missed Participation: Participation is open for two weeks at a time and does not require a large time commitment to receive full marks. Therefore, no accommodations will be made for missed participation marks.

Missed Assignments or Mini Project: There are no accommodations for missed assignments or mini project. Extensions may be granted in extreme situations at the discretion of the instructor if received no later than 24 hours prior to the deadline. The mini project must be submitted as part of the minimum work requirement.

Missed Term Test:

- If a student missed the term test for a valid medical reason, and has **both** filled out the absence declaration form on ACORN **and** emailed the instructor with details and confirmation of completion of the absence declaration **within one week** of the test, then they have the opportunity to write a make-up test at a date and time scheduled by the instructor.
- If the test is missed for any other reason, prior approval must be obtained from the instructor in order to be eligible to write the make-up test. If approval to miss the term test has not been obtained before the test takes place, a grade of zero will be given.
- To meet the minimum work requirement for this course, you must write the make-up if you missed the term test for a valid and documented medical reason, otherwise a meaningful grade cannot be calculated for you to pass the course.
- NOTE: since the term test is online and all students receive the test link, if you write the test or look at the test questions, you forfeit your eligibility to write the make-up test. Therefore you must make your decision about whether you are well enough to write the test before the test has begun.

Missed Final Assessment: The final assessment must be completed in order to meet the minimum work requirement to pass the course, so no accommodations will be made for missing the final assessment. Student will be given ample time to complete the assessment and extensions in general will not be granted.

REGRADE REQUESTS

Regrades will be accepted for the term test and the mini project. Regrade requests must provide a justification for where there exists a grading error and/or how the work meets the grading rubric. All regrade requests will be accepted through an automated process (details provided later) and will be accepted no later than one week after the grades are released. No regrade requests will be accepted by email.

INTELLECTUAL PROPERTY

Course materials provided on Quercus, such as lecture slides, assignments, tests and solutions are the intellectual property of your instructor and are for the use of students currently enrolled in this course only. **Providing course materials to any person or company outside of the course is unauthorized use**. This includes providing materials to predatory tutoring companies.

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:

- You may consult class notes/lecture slides during tests and projects, 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.
- Paying anyone else to complete your assessments for you is academic misconduct.
- Sharing your answers/work/code with others is academic misconduct.
- Looking up solutions to test problems online or in textbooks and copying what you find 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 from anyone or anywhere else. Unacknowledged copying or unauthorized 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 dont plagiarize.

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.

CLASS SCHEDULE - TENTATIVE

Below is a tentative schedule and list of topics to be covered in class, corresponding to Chapters 1-7 in Sheather, with occasional review from other courses as needed. The schedule is subject to change.

Week	Content	Textbook
1	**Introduction:** syllabus, motivating example(s), review of mathematical/statistical concepts needed, introduction to R/RStudio	Ch. 1
2	**Simple linear regression:** Model and Least Squares approach for parameter estimation, variance of error term, assumptions	Ch. 2.1
3	**Inference in Simple Linear Regression Part 1:** review of relationship between Z and T distributions and confidence interval theory, inference on the slope and intercept, confidence intervals for population regression line	Ch. 2.2-3, 2.7
4	**Inference in Simple Linear Regression Part 2:** prediction intervals for response, ANOVA and sums of squares, coefficient of determination, using indicator variables in SLR	Ch. 2.4-5, 2.7
Feb. 7	Mini Project Part 1 due by 11:59PM EST on Quercus Discussion # 1 due by 11:59PM EST on Quercus	
5	**Diagnostics for Simple Linear Regression:** residuals and residual plots, leverage and influential points	Ch.3.1-2
Feb. 14	Assignment # 1 due by 11:59PM EST on Crowdmark	
	READING WEEK	
6	**Handling violations in Simple Linear Regression:** transformations to stabilize variance, transformations for non-linearity, Box-Cox	Ch.3.3
Feb. 28	Discussion # 2 due by 11:59PM EST on Quercus	
7	**Weighted Least Squares in Simple Linear Regression:** parameter estimates with weights, using least squares for weighted least squares, residuals	Ch. 4
Mar. 7	Mini Project Part 2 due by 11:59PM EST on Quercus	
8	**Multiple linear regression:** motivation through polynomial regression, review of matrix linear algebra, parameter estimation in MLR, properties of least squares estimates	Ch. 5.1-2
Mar. 9/11	Term Test during scheduled class time	
Mar. 14	Discussion # 3 due by 11:59PM EST on Quercus	
9	**ANOVA and ANCOVA:** Confidence intervals for parameters, F-test, partial F-test, working with indicator/dummy variables	Ch. 5.2-3
10	**Diagnostics for Multiple Linear Regression:** residuals and their properties, standardized residuals, leverage points, residual plots (omit 6.1.3), influential observations	Ch. 6.1
Mar. 28	Assignment # 2 due by 11:59PM EST on Crowdmark Discussion # 4 due by 11:59PM EST on Quercus	

11	**Handling violations and Variable Selection:** transformations, multicollinearity and variance inflation factors, adjusted R-squared, AIC/BIC, Mallows Cp	Ch. 6.2, 7
12	**Variable selection:** variable selection procedures, model validation	Ch. 7
Apr. 9	Assignment # 3 due by 11:59PM EST on Crowdmark Discussion # 5 due by 11:59PM EST on Quercus	
Apr 13-23	Final assessment period - Final project due between April 13-23	