Methods of Data Analysis 1

University of Toronto
Department of Statistical Sciences
STA302H1S/1001HS

Instructor:	Katherine Daignault	**Class Day/Time:**	
Email:	katherine.daignault@mail.utoronto.ca	LEC 0101:	T 10-12 and R 10-11
Office:	HS 376 (155 College St.)	LEC 0201:	T 2-3 and R 1-3
Office Hours:	T 12-1, R 11:30-12:30	**Class Location**	
Course webpages:	Quercus and Piazza	LEC 0101:	HS 610
		LEC 0201:	ES 1050

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, geometry of 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.

Learning Outcomes: By the end of this course, all students should have a solid understanding of linear regression analysis theory, as well as have developed practical skills for modelling data with linear regression and interpreting corresponding results.

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: All lecture slides and materials will be posted on the Quercus course page. Further, any important announcements will also be posted in Quercus. Please make sure to check it regularly so you don't miss anything.

Textbook: We will be following *A Modern Approach to Regression with R* by 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/.

Calculators: You will need a basic (non-programmable, non-graphing) calculator for tests and exams. Phone calculators or any other devices that permit communication or access to Wifi are **NOT** allowed during tests or exams.

Statistical Software: We will be using R or RStudio (the free version) for performing statistical analyses. R is available for download at http://cran.r-project.org/ for Windows, Mac and Linux operating systems. RStudio is a good integrated development environment to R and can also be downloaded for free at https://www.rstudio.com/products/rstudio/download/. Support for downloading and learning R

(and RStudio) will be provided during lectures or through documents on Quercus. In lectures, examples with R syntax will be provided, which should be sufficient for you to do your assignments.

COURSE COMPONENTS

Lectures: During lectures, we will cover important course materials, as well as cover a number of examples illustrating the uses of these methods. Lecture slides will contain some R code and output to show how to perform these methods in practice. Each week builds on the material from previous weeks, so it is recommended that you attend lectures regularly.

Office Hours: Each TA will hold regular weekly office hours in HS 381. The office hour schedule will be posted on Quercus when finalized. In addition, the instructor will hold weekly office hours. It is recommended that you visit office hours whenever you have a question about the material or assignments. It is best not to leave your questions to the last minute.

Piazza: We will be using Piazza as an online discussion forum. **All questions about course material and exams should be posted here** or asked during TA/instructor office hours. Posts on Piazza can be done anonymously. The instructor and TAs will monitor Piazza and will help answer questions but students are encouraged to answer posts and help their fellow classmates. **The instructor will only respond to emails of a private or sensitive nature.** When emailing the instructor, please use your official mail.utoronto.ca email and include your full name and student number. Also include "STA302" in the subject line.

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.

Undergraduate students will be evaluated in the following way:

Assessment	**Date Due/Occurring**	**Scheme 1**	**Scheme 2**
Class Participation	Ongoing	5%	0%
Assignment # 1	Sunday February 2 by 23:59	5%	5%
Assignment # 2	Sunday February 16 by 23:59	7%	7%
Midterm (LEC 0101)	Tuesday February 25	25%	30%
Midterm (LEC 0201)	Thursday February 27		
Assignment # 3	Sunday March 22 by 23:59	7%	7%
Assignment # 4	Friday April 3 by 23:59	6%	6%
Final Exam	Scheduled by FAS	45%	45%

Please note that the last day to drop the course without penalty is March 15, 2019.

Graduate students will be evaluated in the following way:

Assessment	Date Due/Occurring	Scheme 1	Scheme 2
Class Participation	Ongoing	5%	0%
Assignment # 1	Sunday February 2 by 23:59	6%	6%
Assignment # 2	Sunday February 16 by 23:59	8%	8%
Midterm (LEC 0101)	Tuesday February 25	25%	30%
Midterm (LEC 0201)	Thursday February 27		
Assignment # 3	Sunday March 22 by 23:59	9%	9%
Assignment # 4	Friday April 3 by 23:59	7%	7%
Final Exam	Scheduled by FAS	40%	40%

EVALUATION BREAKDOWN

Class Participation: During lectures, I will be posing some questions to you through a software called Poll Everywhere (https://PollEv.com/katherinedai702/register) to see if you are following along.

- You do not need to answer the questions correctly in order for you to receive participation marks, but it would be helpful to me and you if you try to figure out the right answer, so we can both see how comfortable you are with the material. **Participation is optional** - if you choose not to participate, your final grade will come from Scheme 2 only.
- To use the software and get credit for participating, you will need to register using your official **University of Toronto email address**. (more details during the first week of class)

Assignments: There will be 4 assignments throughout the term. They will generally consist of small data analysis projects, but may also contain some more mathematical work as well.

- Assignments can be done in groups of two, and only one copy needs to be submitted per group. It is important that **both students in the group** attempt all questions on the assignment, since you will need to understand how to answer these types of questions for the midterm and final exam.
- Since you are allowed to work in pairs, it will not be appropriate to post questions about the assignments on Piazza. However, generic lecture content or R questions can be posted on Piazza.
- We will be using Crowdmark to collect and grade the assignments - **you will need to upload each assignment to the Crowdmark system by 11:59PM (i.e. 23:59) on the date it is due** - instructions on how to do this will be given with each assignment, and will also be posted on Quercus.
- There will be a 20% penalty for each day that an assignment is late. Late submissions will not be accepted 48 hours past the due date.
- Since you will be given more than a week to complete the assignments, there will be no additional accommodations for late or missed assignments.

Midterm exam: The midterm will be held **during the regular scheduled lecture time** during Week 7. Exact details (including location) will be communicated through Quercus as soon as they are finalized. It will be approximately 1.5 hours long and will cover all lecture materials from Week 1 up to and including Week 6.

Final exam: The final exam will be a 3 hour cumulative exam and will occur during the exam period (April 6-25). The date of the final will be communicated through Quercus as soon as it is scheduled.

MIDTERM/FINAL EXAM INFORMATION

Aids: Both midterm and final exam are closed-book, however you will be allowed a one single-sided 8.5×11 inch handwritten aid sheet for the midterm, and one double-sided 8.5×11 inch handwritten aid sheet for the final exam. These are useful studying tools, so it is recommended that students spend some time on creating and modifying these throughout the term.

Grading/Regrading: Assignments and tests will be graded using Crowdmark. This will allow students to obtain feedback quicker and without the need to hand back paper copies. Regrading requests will only be considered for the midterm and must be made in writing within one week of the grade being released. All regrade requests must provide a justification in order to be considered.

Missed Midterm: There are no make-up tests. If the midterm is missed for a valid reason, the weight of the assessment will be moved to the final exam. Such situations include:

- a valid medical reason: the student must submit a University of Toronto Verification of Student Illness or Injury form to the instructor in person within one week of the missed test. The form will only be accepted if it is the original form, completed as per the instructions, and indicates the degree of incapacitation on academic functioning. Forms indicating a negligible or mild degree of incapacitation will not be considered a valid medical reason.
- other valid reason (e.g. death in the family): the student must obtain prior approval to miss the midterm from the instructor, with supporting documentation as applicable.
- If no valid reason is provided, or prior approval of absence is not obtained, the midterm will receive a grade of 0.

COMMUNICATION

Please do not email the instructor with questions related to the content of the course. It is much easier to have these types of questions answered by coming to either the instructor or TA office hours. Alternatively, you can post them on Piazza and see if your fellow students can help answer your question. Email is appropriate if you have a personal or emergency matter to discuss. If you have to email the instructor, please include "STA221" in your subject line, your student number in the body of the email, and send it from your Utoronto 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**.

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 on assignments:

- It is acceptable to discuss assignment problems with other students in the class as long as answered are not shared between students (other than the student with whom you are working with).
- Do not let other students read your completed assignment solutions as this can lead to copying.
- It is acceptable to get help with your assignments from someone outside the class, but the help must be limited to general discussion and examples that are not the same as the assignments. As soon as you get an outside person to actually start working on one of your assignments, you have committed an academic offence!
- All work that you submit in assignments 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 UT 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. The content corresponds to Chapters 1-7 in the textbook, with occasional review from other courses as needed. The instructor reserves the right to modify this schedule as needed due to time constraints.

Week	Content	Textbook
1	**Introduction:** syllabus, motivating example(s), review of mathematical/statistical concepts needed, introduction to R/RStudio/RMarkdown	Chapter 1
2	**Simple linear regression:** Model and Least Squares approach for parameter estimation, variance of error term, assumptions	Chapter 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	Chapter 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 (Assignment 1 due Sunday at 11:59PM)	Chapter 2.4-5, 2.7
5	**Diagnostics for Simple Linear Regression:** residuals and residual plots, leverage and influential points	Chapter 3.1-2
6	**Handling violations in Simple Linear Regression:** transformations to stabilize variance, transformations for non-linearity, Box-Cox (Assignment 2 due Sunday at 11:59PM)	Chapter 3.3
READING BREAK		
7	**Weighted Least Squares in Simple Linear Regression:** parameter estimates with weights, using least squares for weighted least squares, residuals (Midterm in class)	Chapter 4
8	**Multiple linear regression:** motivation through polynomial regression, review of matrix linear algebra, parameter estimation in MLR, properties of least squares estimates	Chapter 5.1-2
9	**ANOVA and ANCOVA:** Confidence intervals for parameters, F-test, partial F-test, working with indicator/dummy variables	Chapter 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 (Assignment 3 due Sunday at 11:59PM)	Chapter 6.1
11	**Handling violations and Variable Selection:** transformations, multicollinearity and variance inflation factors, adjusted R-squared, AIC/BIC, Mallows Cp	Chapter 6.2, 7
12	**Variable selection:** variable selection procedures, model validation, LASSO (Assignment 4 due Friday at 11:59PM)	Chapter 7