Methods of Data Analysis 1

University of Toronto
Department of Statistical Sciences
STA302H1F/1001HF Summer 2020

	LEC 0101:	**LEC 5101:**
Instructor:	Katherine Daignault	Mohammad Kaviul Anam Khan
Email:	katherine.daignault@mail.utoronto.ca	kaviul.utoronto@gmail.com
Class Day/Time:	TR 2-5PM EST	TR 6-9PM EST
Office hours:	TR 2-4PM EST on Bb Collaborate	TR 5-6PM EST

---

* ***This is an online course. Please note that since lectures and/or evaluations will be taking place during the above lecture times, you must be available during those times. No accommodations will be made for assessments missed during these times.***

** ***As this is an online course and all assessments must be submitted through Quercus, it is the STUDENT'S responsibility to ensure they have a reliable internet connection.***

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, 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: All lecture slides, recordings and materials will be posted on the Quercus course page for each lecture section. 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 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/.

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 RCloud which only requires you to login with your Utoronto email and connect to our course project via the link provided. Support for downloading and learning R (and RStudio/RCloud) 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 learn how to apply the statistical methods.

COURSE COMPONENTS

Lectures: Lectures will take place in one of two online formats: for LEC0101, lectures will be held via pre-recorded lecture videos posted on Quercus; for LEC5101, lectures will be held live on Bb Collaborate through Quercus with recordings posted afterwards. During lectures and vidoes, we will cover important course materials, as well as cover a number of examples illustrating the uses of these methods. Lecture slides/videos will contain some R code and output to show how to perform these methods in practice. Each lecture builds on the material from previous weeks, so it is recommended that you attend lectures regularly/keep on top of the material.

Office Hours: Instructors and TAs will hold office hours through Bb Collaborate in the Quercus course page. The office hour schedule will be posted on Quercus. 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 accelerated 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.

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**
Discussion Board Participation	By Sunday each week	10%	5%
"Weekly" Online Quizzes (×10)	Last 30min of each lecture period	40%	40%
Mini Project # 1	May 24 at 11:59PM EST	15%	15%
Mini Project # 2	June 7 by 11:59PM EST	15%	15%
Final Project	TBD (June 17-25)	20%	25%

Please note that the last day to drop the course without penalty is June 1, 2020.

Graduate students will be evaluated in the following way:

Assessment	**Date Due/Occurring**	**Scheme 1**	**Scheme 2**
Discussion Board Participation	By Sunday each week	10%	5%
"Weekly" Online Quizzes (×10)	Last 30min of each lecture period	30%	30%
Mini Project # 1	May 24 at 11:59PM EST	20%	20%
Mini Project # 2	June 7 by 11:59PM EST	20%	20%
Final Project	TBD (June 17-25)	20%	25%

EVALUATION BREAKDOWN

Discussion Board Participation: Participation is mandatory and will be done through the use of the Quercus discussion board. 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.
- **Graded participation discussion:** Each week we will post a discussion topic based on content presented in the week's lectures. 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 May 11 and participation is mandatory. Topics will remain open for contribution until Sunday night (local time) so it's best not to wait until the last minute to contribute. A rubric will be posted explaining how this will be graded.

"Weekly" Online Quizzes: There will be 10 "weekly" online quizzes, that will be occurring during the last 30 minutes of the lecture time of each section. Quizzes will begin on **Tuesday May 12** and continue until the last lecture period.

- We will take the best 8 quiz marks and drop the worst 2 in the calculation of your overall quiz mark
- The quizzes will be multiple choice and cover material from the previous set of lectures. You may wish to have a calculator available at this time to aid in any calculations.
- Quizzes can be found under Quercus Quizzes in the navigation bar, or through the link provided in that week's module, and will only be available during the designated quiz time. Quizzes must be done individually.
- **Missed quiz:** Because only the best 8 quiz marks will be counted, we will not be making any accommodations for missed quizzes. These will receive a mark of 0, but will be dropped as part of the two worst quiz marks. Therefore, you may miss two quizzes without penalty.
- **There are no make-up quizzes.** Quizzes, beyond the 2 that will be dropped, will be given zero.

Mini Projects: You will be given two mini projects in the term. The purpose of these mini projects is to develop your data analysis skills which will be useful for the final project and future courses. The mini projects 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 projects will be as follows:

- You will be asked to use the methods taught in lecture to perform a small data analysis
- To submit your results, 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 non-statisticians.
- 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.
- In general, extensions will not be given unless a valid reason is provided. In such cases, the instructor may decide to grant an extension of up to 5 days.
- **There are no make-up mini projects**. A missed mini project will be given a grade of 0.

Final Project: The final project will be due between **June 17-25, 2020** 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 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
- Exploratory data analysis section: a detailed description of the variables in the data with appropriate tables or figures that highlight certain characteristics deemed relevant or important.

- Model development 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
- 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 stated deadline. A word count limit will be given, as well as other more detailed instructions at a later date. **In order to pass the course, you must submit the final project.**

MISSED ASSESSMENT POLICY

Students are responsible for completing all of the assessments detailed in the previous section. If a student is sick and needs to request an extension or accommodation on a mini project, they must send an email to their instructor. In order for the request to be considered, the email:

- must be received at least one day before the mini project is due
- must include the course code in the subject line
- must include your full name and student number
- must specify for which project the extension/accommodation is being requested
- must include the following sentences:
  - "I affirm that I am experiencing an illness or personal emergency and I understand that to falsely claim so is an offence under the Code of Behaviour on Academic Matters."
  - "I understand that the weight of this assessment will be moved to the weekly quizzes (10%) and to the final project (5%)"

In order to pass this course, students must submit the final project, at least one mini project and have passed 5 of the quizzes.

COMMUNICATION

Please do not email the instructor with questions related to the content of the course. These types of questions are much easier to answer through the discussion board or during office hours. Emails that do not contain sensitive or personal information will be directed to post the questions on the discussion board. If you need to email the instructor for personal reasons, please use your official University of Toronto email address, include STA302 in the subject and also include your full name and UTORid in the body of the email (in case we need to look anything up).

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 quizzes 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 quiz 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. 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
1a (May 5)	**Introduction:** syllabus, motivating example(s), review of mathematical/statistical concepts needed, introduction to R/RStudio	Chapter 1
1b (May 7)	**Simple linear regression:** Model and Least Squares approach for parameter estimation, variance of error term, assumptions	Chapter 2.1
2a (May 12) Quiz 1	**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
2b (May 14) Quiz 2	**Inference in Simple Linear Regression Part 2:** prediction intervals for response, ANOVA and sums of squares, coefficient of determination, using indicator variables in SLR	Chapter 2.4-5, 2.7
3a (May 19) Quiz 3	**Diagnostics for Simple Linear Regression:** residuals and residual plots, leverage and influential points	Chapter 3.1-2
3b (May 21) Quiz 4	**Handling violations in Simple Linear Regression:** transformations to stabilize variance, transformations for non-linearity, Box-Cox	Chapter 3.3
May 24	Mini Project #1 due by 11:59PM EST on Quercus	
4a (May 26) Quiz 5	**Weighted Least Squares in Simple Linear Regression:** parameter estimates with weights, using least squares for weighted least squares, residuals	Chapter 4
4b (May 28) Quiz 6	**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
June 1	Deadline to drop course without penalty	
5a (June 2) Quiz 7	**ANOVA and ANCOVA:** Confidence intervals for parameters, F-test, partial F-test, working with indicator/dummy variables	Chapter 5.2-3
5b (June 4) Quiz 8	**Diagnostics for Multiple Linear Regression:** residuals and their properties, standardized residuals, leverage points, residual plots (omit 6.1.3), influential observations	Chapter 6.1
June 7	Mini Project #2 due by 11:59PM EST on Quercus	
6a (June 9) Quiz 9	**Handling violations and Variable Selection:** transformations, multicollinearity and variance inflation factors, adjusted R-squared, AIC/BIC, Mallows Cp	Chapter 6.2, 7
6b (June 11) Quiz 10	**Variable selection:** variable selection procedures, model validation, LASSO	Chapter 7
June 17-25	Final assessment period - Final project due between June 17-25	