Methods of Data Analysis I

University of Toronto
Department of Statistical Sciences
STA302H1F Summer 2022
Instructor: Mohammad Kaviul Anam Khan
Email: sta302@utoronto.ca
Class Day/Time: MW 2pm - 5pm ET at SF1105
Office hours: MW 5pm - 6pm ET at SF1105

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, penalized regression, Generalized Additive Models (GAM) and principal component analysis (PCA). 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 tool. 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.

Textbooks: We will be mostly 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/`. However, some of the lecture materials will be covered from some other books such as*Linear Regression Analysis* by Douglas C. Montgomery et.al (Wiley) and *Generalized Additive Models: An Introduction with R* by Simon N. Wood (CRC press, 2017). However, these books are optional and you don't need to buy. Rather the focus should be on the lecture materials.

Statistical Software: We will be using R with 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 University of Toronto JupyterHub (**link**) with RStudio selected which only requires you to login with your Utoronto email and connect to our course project via the link provided. 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 be held live in-person at SF1105. 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 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: Instructor office hours will be held right after the lectures are concluded in the same room.The TAs will hold office hours through zoom starting week 2. 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. Please 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

All the students will be evaluated in the following way:

Assignment 1	Will be assigned on 18th May and due on June 1st at 11:59PM ET	15%
Assignment 2	Will be assigned on 1st June and due on June 15th at 11:59PM ET	15%
Term Test	May 25th in class	25%
Final Exam	June 22-27	45%

Please note that the last day to drop the course without penalty is June 6, 2022

EVALUATION BREAKDOWN

Assignment: You will be given two assignments in the term. The purpose of these assignments is to develop your data analysis skills which will be useful for developing data analysis skills as well as to develop practical understanding of the methods taught in the class. The assignment will have a heavy focus on the use of statistical software (R specifically), and will involve applying the methods learned during lecture to a data set. The format of the assignments will be as follows:

1. use the methods taught in lecture to perform a small data analysis.
2. simulate unique datasets and writing your own functions instead of built in R functions.
3. solve some mathematical problems and explain the procedure
4. In general, extensions will not be given unless a valid reason is provided. Please let the instructor know about the reasoning within 48 hours of the submission deadline. In such cases, the instructor may decide to grant an extension of up to 2 days.
5. **There are no make-up for assignments**. A missed assignment will be given a grade of 0.

Term Test: The term test will be held on May 25th in class. More details will be provided during the second week of lectures.

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.

In order to pass this course, students must complete the final exam and at least 1 of the assignments.

MISSED ASSESSMENT POLICY FOR THE TERM TEST

Students are responsible to attend the assessments. If a student is sick and needs to request a re-weighting for the assessment. In order for the request to be considered, the email:

- must be received within 48 hours after assessment is due
- must include the course code in the subject line
- must include your full name and student number
- 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 offense under the Code of Behaviour on Academic Matters."
  - "I understand that the weight of this assessment (term test) will be moved to the assignments (10%) and to the final exam (15%)"

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:

- Students must complete all assessments individually. Working together is not allowed.
- Having anyone else to complete your assessments for you is academic misconduct.
- Sharing answers/work/code for STA302 assessments with any other student is academic misconduct.
- Looking up solutions to assessments problems online or in textbooks and copying any part of what you find is an academic offense.

- 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 or must not have possession/use of unauthorized aids or assistance associated with tests during the tests. 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 University of Toronto Policy on Cheating and Plagiarism, and don't plagiarize.
- will

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

Week	Content
1a (May 9)	**Introduction & Inference in Simple linear regression Part 1:** syllabus, motivating example(s), review of mathematical/statistical concepts needed, introduction to R/RStudio. Linear Model and Least Squares approach for parameter estimation, error variance and confidence interval theory.
1b (May 11)	**Inference in Simple Linear Regression Part 2:** Inference on the slope and intercept, confidence intervals and prediction intervals, ANOVA , coefficient of determination, indicator variables.
2a (May 16)	**Diagnostics for Simple Linear Regression:** residuals and residual plots, leverage and influential points .
2b (May 18)	**Handling violations in Simple Linear Regression:** transformations to stabilize variance, transformations for non-linearity, Box-Cox
May 23	University Holiday (no class or office hours)
3b (May 25)	Term Test.
3a (May 30)	**Multiple linear regression:** motivation through polynomial regression, review of matrix linear algebra, parameter estimation in MLR, properties of least squares estimates.
4a (June 1)	**ANOVA and ANCOVA:** Confidence intervals for parameters, F-test, partial F-test, working with indicator/dummy variables
4b (June 6)	**Diagnostics for Multiple Linear Regression:** residuals and their properties, standardized residuals, leverage points, residual plots, influential observations. Assignment 1 due.
5a (June 8)	Weighted and Generalized Least Squares in Multiple Linear Regression Handling violations and Variable Selection.
June 6	Deadline to drop course without penalty
5b (June 13)	**Variable selection:** variable selection procedures, model validation, Ridge regression and LASSO
6a (June 15)	**Data Reduction:** Principal Components Analysis. Orthogonal Variables
6b (June 20)	Generalized Additive Models. Assignment 2 due.
Final Exam	Final assessment period (June 22-29).