STA302/STA1001: Methods of Data Analysis I Summer 2023

Nnenna Asidianya

E-mail: nnenna.asidianya@mail.utoronto.ca
Office Hours: TBA

Class Hours: Mon 6:00pm-9:00pm
Wed 6:00pm-9:00pm (in MB128)

- ***This is in person 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.***
- ***Any assessments that take place outside of the class setting will be submitted via Crowdmark.***

Course Description

This course covers theory and applications of regression analysis. We will develop the theory of regression models and study how to analyse data when such models are appropriate. Topics to be covered include: simple and multiple regression models using least squares, analysis of variance, inference for regression parameters when the errors are normally distributed, confidence and prediction intervals, geometry of least squares, multicollinearity, regression models for quantitative and qualitative predictors, model selection and validation, and diagnostics.

Prerequisites/Corequisites

Pre-requisites: 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. **If you can advise students to contact our undergraduate office at ug.statistics@utoronto.ca if they have questions about course prerequisities.**

Email Policy

Should you have any regrade concerns to relay to the teaching team, please direct your email to questions to sta302@utoronto.ca. Since there may be a high volume of concerns at a given time, please allow for a 48 hour reply policy. **In the email make sure to state your name and student**

number as it appears on Quercus in the body of the email.

For highly personal concerns please refer to the following :

- For serious concerns email me at nnenna.asidianya@mail.utoronto.ca
- Use your academic email, for example mine is nnenna.asidianya@mail.utoronto.ca
- Use the following format in the subject of your email: CourseName/LASNTNAME For example: STA302S/ASIDIANYA
- Be very clear and concise.

Required Materials

- This course requires the following textbooks:
  1. **Kutner, M. H., Nachtstein, C. J., and Neter, J. Applied Linear Regression Models. McGraw-Hill, 5th edition.**
  2. **Sheather, S.J. A Modern Approach to Regression with R. (Springer).**
- 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 R framework
  2. RStudio is a good integrated development environment to R (makes it simpler to work in R) and can also be downloaded for free at RStudio.

  For students who are unable to download R, they may be able to use the following local cloud: Jupyter. You will need to use your UTORID and password.

Course Components

Class Lectures: The lectures will take place in person, in MB128. There are two three hours a week class times will be used to cover important course materials. It is important that you attend these classes in order to keep up with the topics, and gain a deeper understanding of the theories and applications of statistics. It is beneficial to bring your laptop to follow along with code.

Office Hours: The instructor and TAs will hold office hours through a combinaion of zoom and in person. The office hour schedule will be posted on Quercus when it is finalized. It is recommended that you visit office hours whenever you have a question about the material. This is very important since this is an online accelerated class, so onus is up to the student to have material clarified as quickly as possible. Don't wait until the last minute to ask your questions!

Piazza board: This will be used as an online communication board. Any questions related to course materials should be posted there. If you have any administrative questions this should be directed to me via email (see the Email Policy above).

Course Assessment

Assignments: Over the course of the semester students will be given a mini assignment to work on that is related to the course material up until that point. This may include both an R component and a written component. More details, such as the content and deadline, will be communicated later. **No late report will be accepted**.

Undergraduate students will be evaluated in the following way:

Item	Date Due	% of Grade
Term Test	July 26th (in Class)	25%
Mini-Assignment	Due July 19th by 5PM EST	15%
Final Project	Due August 14th by 5:00PM EST	30%
Final Exam	To Be Announced	30%

Term Test

The term test will take place in class time on July 26th, from 6-8PM EST. The term test will cover all the material covered in class and exercises until the last week before the test. Information about the break down of the test will be released closer to the exam date.

Mini Assignment

You will be given one mini assignment in the term. The purpose of this mini assignment is to develop your skills which will be useful for the final project at the end of the term. The mini assignment will have a heavy focus on the use of statistical software (R ally), and will involve applying the theoretical and applied methods discussed in class.

Final Project

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 an appropriate data set to perform linear regression, and applying the correct methods to apply and providing appropriate justifications where necessary. **This is a formal report and therefore it must contain the following sections:**

- Introduction section: provides details regarding the question you wish to address, why the model is being developed, how you intend to go about developing the model, and finally how the model meets the purpose mentioned earlier.
- Exploratory data analysis section: a detailed description of the variables in the data set with appropriate tables or figures that highlight certain characteristics of your variables that you deem important to mention. This will be important for contextualizing why linear regression makes sense for your model.

- Model development section: a detailed discussion of the process used to come to the final model. Justifications may be both statistical and empirical in nature.You should also have 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 (i.e explain how the variables work, discuss predictions), and discuss any limitations/problems remaining with the model and how they might impact its use in the real world.

The timelines for the project are as follows:

Step	Description	Due Date	% of Grade
1	Complete the Class Survey	Aug. 4th, 2021, 5:00 PM in Quercus	2%
2	Final Report	August 15th, 5:00PM EST	23%

Missed Assessment Policy

Students are responsible for completing all of the assessments detailed in the previous section. In general, there is a 10% deduction per day up until day 5 for late or missed work. If a student is sick and needs to request an extension or accommodation on an assessment, they must send an email to their instructor. In order for the request to be considered, the email:

- for a missed mini assignment, notification about an extension must be received **at least** 48 hours before the mini project or term test is is due
- for an extension pertaining to the final project, notification of the accommodation or extension must be provided within 48 hours. **You may receive an extension of up to 72 hours at most.**
- the subject line must be written in the format shown in *Email Policy*
- must include your full name and student number in the body of the email
- 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."

Remark Policy

Any requests to have marked work re-evaluated must be made in writing to the email: **sta302@utoronto.ca within one week of the date the work was returned to the class**. The request must contain a **valid** justification for consideration. You are responsible to check that your scores are entered correctly on Quercus. Any requests for a mark that was not entered correclty in Quercus must be made in writing within one week of the date the mark was entered in Quercus.

"Note that your entire assessment may be remarked and your assessment grade may remain the same, go up, or go down."

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.

What is not permitted is providing materials to predatory tutoring companies, or to friends who are not officially enrolled in this course this term.

Providing course materials to any person or company outside of the course is unauthorized use. This includes providing materials to predatory tutoring companies.

Accessibility Statement

Students with diverse learning styles and needs are welcome in this course. 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

Academic Integrity Statement

Academic integrity is essential to the pursuit of learning and scholarship in a university, and to ensuring that a degree from the University of Toronto is a strong signal of each student's individual academic achievement. As a result, the University treats cases of cheating and plagiarism very seriously. The University of Toronto's Code of Behaviour on Academic Matters (http://www.governingcouncil.utoronto.ca/policies/behaveac.htm) outlines the behaviours that constitute academic dishonesty and the processes for addressing academic offences. Potential offences include, but are not limited to:

IN PAPERS AND ASSIGNMENTS: Using someone else's ideas or words without appropriate acknowledgement. Submitting your own work in more than one course without the permission of the instructor. Making up sources or facts. Obtaining or providing unauthorized assistance on any assignment.

ON TESTS AND EXAMS:Using or possessing unauthorized aids. Sharing, posting, or discussing questions or answers with anyone in or outside the course. Misrepresenting your identity.

IN ACADEMIC WORK: Falsifying institutional documents or grades. Falsifying or altering any documentation required by the University, including (but not limited to) doctor's notes. All suspected cases of academic dishonesty will be investigated following procedures outlined in the Code of Behaviour on Academic Matters. 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/).

Schedule and weekly learning goals

Table 1: Tenative Weekly Schedule

Week 1	Relation between variables (p.2-3) Regression models and their uses (p.5-9) Data for regression analysis (p.12-13) Overview of steps in regression analysis (p.13-15)	Introduction
Week 1 cont.	Model and assumptions (p.9-12) Least squares estimation (p.15-22) Predicted values, residuals (p.22-23)	Simple linear regression I
Week cont.	Analysis of variable in the dependent variable (p.63-68) Normal errors regression model, maximum likelihood estimation (p.2632) Precision of estimates (p.40-63)	Simple linear regression II
Week 2	Inference (p.40-63) Gauss-Markov theorem (p.18-19) Regression through the origin (p.161-165) Matrices and vectors (p.176-197)	Simple linear regression III

Table 2: Tenative Weekly Schedule

Week 2 cont.	Multiple linear regression and matrix approach (p.197-199,214-223) Estimation (p.199-201,223-224) Fitted values and residuals (p.202-204,224-225)	Multiple Linear Regression I
Week 3	Properties of linear functions of random vectors Properties of regression estimates (p.227-232) Normal errors, maximum likelihood estimation	Multiple Linear Regression II
Week 3 cont.	Analysis of variance and quadratic forms (p.204-206,225-227) Hypothesis testing and confidence intervals for single regression parameter, mean response, and prediction of a new observation (p.227-232) Hypothesis testing for several regression parameters: ANOVA and distribution of quadratic forms (p.256-268)	Multiple linear regression III
Week 4	Multicollinearity (p.278-289) The geometry of least squares Violation of assumptions	Multiple linear regression IV

Table 3: Tenative Weekly Schedule

Week 4	cont.	An example with R	Interlude
Week 5		Polynomial regression model (p.294-305) Interactions (p.306-313) Qualitative predictors (p.313-321)	Regression models for quantitative and qualitative predictors
Week 5 cont.		Overview (p.343-350) Analyzing Associations Between Categorical Variables Criteria for model selection (p.353-361) Model validation (p.369-375)	Model selection and validation
Week 6		Model adequacy (p.384-390) Outlying Y observations (p.390-398) Outlying X observations (p.398-400) Influential cases (p.400-406)	Diagnostics