Statistical Sciences
UNIVERSITY OF TORONTO

STA 302H1F / 1001HF 2020
METHODS OF DATA ANALYSIS I / APPLIED REGRESSION ANALYSIS

September 10 - December 23, 2020 (except November 9-13)

Land Acknowledgement

We wish to acknowledge the land on which the University of Toronto operates. For thousands of years it has been the traditional land of the Huron-Wendat, the Seneca, and most recently, the Mississaugas of the Credit River. Today, this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to have the opportunity to work on this land.

Note:	This is a fully online course. Live class sessions will be held via Quercus. Students are responsible for ensuring that they have reliable internet.
Live Sessions:	**SECTION L0101/2001**, Tuesdays 10:10-11:00 ET **SECTION L0201**, Tuesdays 15:10-16:00 ET **SECTION L0301**, Wednesdays 14:10-15:00 ET
Course website:	Available through https://q.utoronto.ca (UofT Quercus)
Teaching Team:	
Instructor:	**Dr. Shivon Sue-Chee** (*she, her, hers*) **E-mail**: shivon.sue.chee@utoronto.ca **Office hours**: Drop-in hours or by appointment (from Sept. 18)
Teaching Assistants:	Names and office times will be posted later in our website.

1 SYLLABUS CONTENT

COURSE OVERVIEW		**pages 2-3**
Course description	Learning Outcomes	Pre-requisites
Requisite and Accreditation status	Required Textbook	Recommended Readings
Computing		

ASSESSMENT AND POLICIES		**pages 4-6**
Assessment	Regrading Policy	Missed Final Exam Policy
Academic Integrity	Intellectual Property	

SUPPORT AND ACCOMMODATIONS		**pages 6-7**
Course website	What to expect during class	Weekly student schedule
Online discussion forum	Accessibility Needs	Communication
Succeeding in our course		

SCHEDULE AND IMPORTANT DATES	**page 8**

2 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. This course will also be an opportunity to develop skills in data analysis for which `R` software and `Rmarkdown` will be used.

Learning Outcomes

By the end of the course, students should be able to do the following, using R/RStudio where appropriate:

- judge when it is appropriate to use regression analysis and setup the appropriate data for linear regression analysis,
- understand the least squares, maximum likelihood and Bayesian approaches for regression parameter estimation,
- build simple and multiple linear regression (SLR and MLR) models, and interpret the regression parameter estimates in non-technical language,
- formulate SLR and MLR models in matrix terms, and mathematically derive the properties of least squares estimates of linear regression parameters under standard assumptions,
- carry out statistical inference for linear regression parameters and the regression line, and interpret the results,
- assess model adequacy using appropriate data visualizations and model outputs,
- identify and apply suitable steps to overcome violations of the assumptions of linear regression,
- understand and apply variable selection methods such as AIC, BIC and cross validation to build valid linear regression models,
- clearly communicate the results of a data analysis to both technical and non-technical audiences, and
- critique data analysis reports or publications which include linear regression models.

Pre-requisites

Pre-requisites will be strictly enforced by the department, not the instructor. Students should have a second year statistics course such as STA238, STA248, STA255, STA261 or ECO 227, a computer science course such as CSC108, CSC120, CSC121, or CSC148 and a mathematics course such as MAT221(70%), MAT223, or MAT240, or equivalent preparation.

You will need to utilize basic matrix operations. The course textbook includes matrix algebra review materials and additional resources will be available in our course website. Your computing experience will be transferred to learning and/or extending our use of `R` (and `RStudio`).

Requisite and Accreditation status

STA 302/1001 is a pre-requisite for many STA courses such as 303/1002, 305/1004 and 414/2104 .

STA 302 is an accredited course under the Canadian Institute of Actuaries (CIA)'s University Accreditation Program (UAP) program. The minimum grade for you to use this course for CIA credentialing purpose is

77. For detailed information on UAP, please visit the following link: CIA University Accreditation

Required Textbook

- Sheather, S. (2009) *A Modern Approach to Regression with R*. New York: Springer. UofT library link at `http://go.utlib.ca/cat/6928488`

  Chapters 1-7 contain relevant materials. Available for purchase at the UT Bookstore and as an electronic resource through the University of Toronto library website. Datasets and other resources are available at the textbook's website: http://gattonweb.uky.edu/sheather/book/.

  Practice problems from the textbook(s) will be posted on the course website. An excellent reference for solutions, done using SAS, can be found at Prof. Gibbs website:

  http://www.utstat.utoronto.ca/alisong/Teaching/1112/Sta302/pp.html

Recommended Readings

- Faraway, J. (2005) *Linear Models with R*, Taylor and Francis. UofT library link at `http://go.utlib.ca/cat/11575808`
- Kutner, M., Nachtsheim, C., and Neter, J. (2004) *Applied Linear Regression Models, 4th edition*. New York: Mc-Graw Hill. Chapters 1- 8 contain relevant materials.
- Weisberg, S. (2009) *Weisberg: Applied linear regression, 4th edition* New York: Wiley). UofT library link to third edition at `http://go.utlib.ca/cat/7834385`
- Wakefield, J. (2013) *Frequentist and Bayesian Regression Methods*. New York: Springer. UofT library link at `http://go.utlib.ca/cat/8824615`

Computing

We will use `R` and `RStudio` for statistical computing. The main advantage of `R` is that it is a freeware and there is a lot of available help resources online. `RStudio` is an integrated development environment (like SAS) to `R`, which makes it easier to work in `R`. `R`/`RStudio` can be downloaded onto your personal computer or used via our university's web server or in the `RStudio` Cloud. If you would use `R` on your personal computer, then installation is via a two-step procedure:

1. Download the base `R` framework at http://cran.r-project.org/ for Windows, Mac and Linux operating systems.
2. Then download `RStudio` for free at https://www.rstudio.com/products/rstudio/download/.

Support for downloading and learning `R` and `RStudio` will be provided by the teaching team (Instructor and TAs). Additional resources will be given in our website. In lectures, examples with `R` syntax will be provided, which should be sufficient for you to practice and do your assignments.

For each assignment, it would be required that you submit a reproducible `RMarkdown` file with your codes and a knitted RMarkdown document as your data analysis report. To learn more about `RMarkdown`, refer to

http://rmarkdown.rstudio.com/index.html

3 ASSESSMENT AND POLICIES

Assessment

Students will be evaluated based on the following scheme.

	Weight For Undergraduates	Weight For Graduates	Date	Time
Quizzes	40%	30%	on most Thursdays	(30 minutes)
Course Surveys	2%	2%	mid & end-of-course	
Participation	8%	8%	by most Saturdays	by 11:59am
Assignment 1	5%	5%	*Sat., Sept. 26	due by 8pm
Assignment 2	10%	15%	*Sat., Oct. 24	due by 8pm
Assignment 3	15%	20%	*Sat., Dec. 5	due by 8pm
Final Exam **	20%	20%	Between Dec. 11-22	(2 hours)

Students must complete the final exam, at least assignment 2 or 3, and five (5) quizzes in order to pass this course.

*Assignments' due dates are subject to change with sufficient notice. Graduate students will be evaluated at the graduate level according to the University Assessment and Grading Practices Policy.

** Graduate students will be required to complete a final project and presentation, instead of a final exam.

Weekly Quizzes There will be ten (10) "weekly" quizzes. Quizzes will be held online in Quercus on Thursdays, beginning September 24.

- Each quiz will be for 30 minutes and will be available from 6am to 12noon on Thursdays.
- The quizzes will have multiple choice and/or short-answer questions, and cover material from the current week's lectures.
- Your best 8 quizzes will count toward your overall quiz grade. Hence, you may miss up to two quizzes without penalty.
- Quizzes must be done individually.
- A scientific calculator may be a useful tool to aid calculations.
- Since the best 8 quizzes will be counted, there will be no accommodations for missed quizzes.

Course Surveys There will be two short course evaluation surveys. One will be held during the mid-course period and the latter will be at the end of the course. More details of these surveys will be announced later.

Participation Activities **Participation is optional** and will take various forms such as hand-written proofs, summaries and peer reviews of assignments. Roughly, participation will be held on a weekly basis and will count during the period September 24 to December 9.

- Each activity will be worth 1% and due by 11:59am on Saturday of each week.
- Participation must be done individually.
- You will be given 1 to 3 days to complete each activity.
- We will use Quercus, Crowdmark and/or peerScholar for these activities.
- If you miss a participation activity, the weight will be shifted to the final exam.

Assignments The assignments will each be a short data analysis project for which you will use `RStudio`. Assignments are to be submitted online into Quercus and/or peerScholar by 8pm on the due dates. Each assignment will consist of two parts: an RMarkdown file and a written report or oral presentation.

- Assignments are short and should take, on average, one week to complete. However, additional time is built into the due date to allow for students with accommodations.
- Late assignments will be accepted but subject to a 20% penalty per day late. Late submissions will not be allowed beyond 48 hours of the due date.
- Students who would like additional accommodations should email the instructor at least one day before the assignment is due.
- Each will ask for an RMarkdown file and the corresponding report. Assignments 2 and 3 will each require a short video presentation.
- There are no make-up assignments. A missed assignment will be given a grade of 0.
- Each assignment can receive two peer reviews.

Final Exam The final exam is cumulative. Further details will be announced later in the course.

Re-grading Policy

Any requests to have marked assignments re-evaluated must be made in writing by email to regradesta302@gmail.com within one week of the date the work was returned to the class. The request must contain a justification for consideration. Be sure to include your official name and student number for identification purposes. Regrading requests should be processed by the teaching team within two weeks of the request date. Please note that the teaching team reserves the right to review a part or the whole of your assignment. Hence, your marks may go down, up or remain the same.

Missed Final Exam Policy

If the final exam is missed, for any reason, a makeup oral exam will be offered within 2 months of the original exam date.

Academic Integrity

The University supports acting in honesty, trust, fairness, respect, responsibility, and courage in all academic matters. Students are responsible for knowing the content of the University's Code of Behaviour on Academic Matters. All suspected cases of academic dishonesty will be investigated following procedures outlined in the Code of Behaviour above. 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 (`http://academicintegrity.utoronto.ca/`). Here are a few guidelines that apply to this course:

- Students may consult course materials during quizzes, however, sharing or discussing questions and answers is an academic offence.
- Instructions for each assessments should be strictly followed. All assessments must be completed individually.
- Do not personate another person, or have another person personate at any assessment.
- It is acceptable to get help with your assignments from someone outside the course, but the help must be limited to general discussion and examples that are not the same as the assignments. As soon as you get some else to actually start working on one of your assignments, you have committed an academic offence!

- You must not copy mathematical derivations, computer output and input, or written descriptions from anyone or anywhere else, without reporting the source within your work. This includes copying from solutions provided to previous semesters of this course. Please read the UofT Policy on Cheating and Plagiarism, and don't plagiarize.

Intellectual Property

Course materials provided on Quercus, such as lecture videos and slides, assignments, quizzes 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.** Failure to comply can result in legal action against all parties involved.

4 SUPPORT AND ACCOMMODATIONS

Course website

The course website is available through Quercus via

https://q.utoronto.ca

and will be used to post video lectures, lecture notes, practice problems, quizzes, assignments, announcements and grades.

What to expect during live class sessions

Live sessions will be conducted via Bb Collaborate mainly by the instructor, with TAs as supporting moderators. These sessions may be recorded for training of the teaching team but will not be made available to students. They are intended to offer a brief overview of the week's materials, and a formal opportunity to discuss the week's materials and interact with other members of the class. To allow for these activities in Bb Collaborate, students are advised to attend their section-specific class session.

A typical 60-minute session, beginning at the 'x'-th hour, will have the following program:

x:00- x:10	Entry and setup
x:10- x:15	Ice breaker poll questions
x:15- x:30	Instructor's overview of week's content
x:30- x:45	Breakout rooms: small group discussion, post questions in chat
x:45- x:55	Round-table discussions
x:55- x:60	Wrap-up

Weekly suggested student schedule

Day	Estimated minimum time	Main student activity
Monday to Wednesday	60m	Learn week's module
Monday to Wednesday	60m	Discuss on Piazza, Group learn, Attend office hours
Tuesday or Wednesday	60m	Attend live class
Thursday	30m	Answer quiz
Friday	60m (optional)	Attend instructor's informal office hour (TBA)
Thursday to Saturday	30m	Complete participation activity

Online discussion forum

This semester you will have the option to use Piazza for class discussion. This is a collective discussion forum for sections- L0101, L2001, L0201 and L0301. The Piazza system is highly catered to getting you course material help fast and efficiently from classmates, the TAs, and myself. Rather than emailing questions, I encourage you to post your questions on Piazza. To sign up for the discussion forum go to the link at:

https://piazza.com/utoronto.ca/fall2020/sta3021001

If you decide not to use Piazza, it will not disadvantage you in any way, and will not affect official University outcomes (e.g., grades and learning opportunities). If you choose not to opt-into Piazza, then you can ask questions or discuss course material with the instructor or TAs during online office hours.

Be sure to read Piazza's Privacy Policy and Terms of Use carefully. Take time to understand and be comfortable with what they say. They provide for substantial sharing and disclosure of your personal information held by Piazza, which affects your privacy. If you decide to participate in Piazza, only provide content that you are comfortable sharing under the terms of the Privacy Policy and Terms of Use.

Note that moderation of the forum is subject to TA availability and further details of how and when the forum will be moderated, will be announced in our website. Please use the forum in accordance with its purpose. Inappropriate posts will not be tolerated and will be dealt with accordingly. **On Thursdays, when quizzes will be held, the forum will be closed temporarily.**

Accessibility Needs

The University of Toronto is committed to accessibility. If you require accommodations for a disability, or have any accessibility concerns about the course, the classroom, or course materials, please contact Accessibility Services as soon as possible at accessibility.services@utoronto.ca or https://www.studentlife.utoronto.ca/as.

Email communication with your instructor

E-mail is appropriate for emergencies or private matters. Use your *utoronto.ca account. You will not get a response if you email from other email addresses. Write a proper email including the course number in the subject line. The email should contain the addressee, your official name and UTORid for identification purposes. Alternatively, you can use the Quercus Inbox via the course website to email me. I will generally answer e-mail within two business days.

Announcements will be posted on Quercus. Please check there regularly. If an urgent matter arises, I may contact the entire class by e-mail. In order to receive these messages, ensure that your account is connected to your *utoronto.ca email and email notifications are enabled.

The TAs and instructor are here to help you. Ask questions and let me know promptly if there are any concerns. TAs will not be available by email.

Succeeding in our course

Students are encouraged be active learners by learning the course materials, staying connected via our course website, being involved in live classroom sessions and diligently completing assessments. Here are some suggestions for achieving success in our course:

- Connect with the instructor and/or TAs during live sessions and office hours.
- Post and answer questions on the discussion forum.
- Join a recognized study group.
- Get advice on learning (outside of the classroom) from a learning strategist.
- Email the instructor in cases of emergencies or personal matters.

5 SCHEDULE AND IMPORTANT DATES

This schedule is subject to change, with adequate notice.

Week	Textbook Section	Topics	Important dates
1	1	**Introduction to data analysis**: types of studies, variable types, data processing, statistical methods, statistical computing, motivating examples	Sept 15/16: Live! Sept 18: Waiting list off
2	2.1	**Simple Linear Regression (SLR)**: the simple linear regression model, least squares, maximum likelihood and bayesian approaches to parameter estimation, SLR in R	Sept 23: Add day Sept 24: Q1 **Sept 26: A1 due**
3	2.2-3, 2.7	**Estimation and Theory for SLR**: review of statistical theory, properties of least squares estimates	Oct 1: Q2
4	2.4-5, 2.7	**Inference in SLR**: confidence and prediction intervals, regression ANOVA,	Oct 8: Q3
5	3.1-2	**Diagnostics in SLR**: using diagnostics plots, outliers, leverage and influential points, standardized residuals, cook's distance	Oct 12: Thanksgiving Oct 15: Q4
6	3.3	**Remedies I**: variance stabilizing transformations, interpreting log-log transforms	Oct 22: Q5 **Oct 24: A2 due**
7	4 (skip 4.1.1, 4.1.2)	**Weighted Least Squares**: parameter estimation with weights, using least squares for weighted least squares smoothing techniques	Oct 29: Q6
8	5.1-2	**Multiple Linear Regression**: polynomial regression, estimation and inference, matrix formulation	Nov 5: Q7
		November 9-13: READING WEEK. No classes	Nov 9: Drop day
9	5.2-3	**ANOVA and ANCOVA**: partial F-test, global F-test, interaction	Nov 19: Q8
10	6.1	**Diagnostics in MLR**: residuals, residual plots, leverage points, standardized residuals, added-variable plots	Nov 26: Q9
11	6.2-6.7, 7 (skip 6.3)	**Remedies II and Variable Selection**: transformations, multicollinearity, variance inflation factors, AIC/BIC, cross validation	Dec 3: Q10 **Dec 5: A3 due**
12	7	**Course Summary and Data ethics**	Dec 9: Last class
		Dec 11-22: Exam period	**TBA: EXAM day**