

Methods of Data Analysis 1

University of Toronto
Department of Statistical Sciences
STA302H1S/1001H1S Winter 2022

Instructor:	Katherine Dagnault	Synchronous Class:	W 5-7PM ET
Course email:	sta302@utoronto.ca	Office Hours:	Th 4-5PM ET
Course webpage:	Quercus	Format:	online using MS Teams

COURSE OVERVIEW

How will this course operate? This course will be offered entirely online, with a combination of synchronous lectures and asynchronous video lectures. Each week, you should view the asynchronous course materials (posted in a Quercus module) and complete the knowledge check quiz. Then come to the synchronous classes which will occur through Microsoft Teams. These sessions will focus on demonstrations of concepts and applications using statistical software, with opportunities for hands-on practice. It is your responsibility to make sure you are available during scheduled lecture times and stay on top of the course material and all relevant deadlines. Note that while the synchronous lecture time is a 2-hour block, we will not usually meet for the full duration but rather the time is available for additional question periods as needed.

**** To avoid any issues accessing materials or lectures, always sign in to your Utoronto Microsoft account and ensure that Microsoft Teams is updated. Avoid using mobile devices to attend lecture or complete assessments.**

***** Please ensure that you have access to reliable internet service, as some assessments will take place online and there is no guarantee that accommodations can be provided for faulty internet during an evaluation.**

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, and model selection and validation. Statistical software will be used throughout and will be required for the completion of various assessments during 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: We have a common Quercus course page for all sections of this course. All lecture slides, recordings and materials will be posted on this Quercus course page. Further, any important announcements will also be posted in Quercus. Please make sure to check it regularly.

Textbook: This course does not strictly follow any particular textbook, but rather merges material from a number of sources. **All of the below recommended textbooks are freely available as an electronic copy through the University of Toronto Library.** Our two primary reference texts will be

- *Linear Models in Statistics*, 2nd edition by Alvin C. Rencher and G. Bruce Schaalje (Wiley).
- *A Modern Approach to Regression with R*, by Simon J. Sheather (Springer)

Other helpful references from which practice problems may be assigned are:

- *Applied Regression Modeling*, 2nd edition, by Iain Pardoe (Wiley).
- *Methods and Applications of Linear Models*, 2nd edition, by Ronald R. Hocking (Wiley)
- *Applied Linear Regression*, 3rd edition, by Sanford Weisberg (Wiley).

These are all useful books, but may present the material in a different order or in a different way. They are still good for additional explanation and practice problems.

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 using RStudio through the [JupyterHub](#) for University of Toronto. This will allow you to login with your official UofT credentials and use RStudio without the need for a local installation. More information about using RStudio in JupyterHub will be provided in the first class. R code shown in class will be available on the course page and, along with any additional resources, should be sufficient to complete any assessment involving data analysis.

COURSE COMPONENTS

Lectures: The majority of the core content for this course will be delivered via pre-recorded video lectures which will be posted to Quercus for each week's module (ideally by Friday night the latest). It is the student's responsibility to watch these videos in a timely fashion. Synchronous classes will occur through Microsoft Teams and will supplement these videos with additional demonstrations and data analyses. Synchronous lectures will also be recorded and posted a few days after the class.

Office Hours: Instructor and TAs will hold office hours online through Microsoft Teams. The office hour schedule will be posted on Quercus once finalized. 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 class to have material clarified as quickly as possible. Don't wait until the last minute to ask your questions!

ED Discussion Board: We will be using the ED-STEM Discussion Board as an online discussion forum, which can be accessed through the Quercus course page. **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.

COMMUNICATION

How your instructor will communicate with you: All communication will be made through Quercus announcements or during lectures. Please ensure that you check Quercus regularly so you don't miss anything important.

Where to send content questions: We will be using the ED Discussion board to collect student questions regarding course content, assignments, etc. All questions should be posted here.

When to email the instructor: The instructor will only respond to emails of a private or sensitive nature. If you email the instructor with content related questions, you will be asked to repost your question on the content board so the answer may benefit all students. Should you need to email the instructor about a sensitive or personal nature, please use your official mail.utoronto.ca email, include your full name and student number in the text. Send all course related emails to sta302@utoronto.ca. Please allow up to 48 hours for a reply. Emails will not be monitored on evenings and weekends.

A note on email and discussion board etiquette: Please make sure that you communicate politely and respectfully with all members of the teaching team and your fellow classmates. Written communications can sometimes take a tone other than what was intended (e.g. can come off as dismissive, rude or insulting), so make sure you re-read or read out loud your email/post before sending it to make sure it has the tone you intended. For more tips on respectful communication, see [professional communication tips](#). The ED discussion board is a teaching and learning tool and therefore should only be used as such. Any posts that detract from the learning goal of the board will be removed to keep the board a safe space.

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 have the grading scheme as outlined below.

Graduate students will use the same grading schemes, with the exception that the Term Test will be worth 15% while the Final Written Report (Part 3) will be worth 25%.

Assessment	Date Due/Occurring	Scheme 1	Scheme 2
Quercus Discussion Participation (4)	Thursdays every 3 weeks	5%	10%
In-class Group Labs (5)	See attached schedule	5%	0%
Weekly Quizzes (10)	Due every Tuesday by 11:59PM ET	15%	15%

Reproducible Writing Exercise (3 parts)			
Part 1: Draft/Create	Jan. 27 by 11:59PM ET	1%	1%
Part 2: Peer Feedback/Assess	Jan. 31 by 11:59PM ET	1%	1%
Part 3: Final Draft/Revise	Feb. 4 by 11:59PM ET	3%	3%
Term Test	Feb. 16 from 5-7PM ET	20%	20%
Video Project	Mar. 18 by 11:59PM ET	20%	20%
Final Project (3 parts)			
Part 1: Research Question/Proposal	Feb. 20 by 11:59PM ET	5%	5%
Part 2: Analysis Flowchart	Apr. 1 by 11:59PM ET	5%	5%
Part 3: Written Final Report	Tentatively Apr. 20	20%	20%

MINIMUM PASSING REQUIREMENT

In order for the instructor to be able to reasonably assess the ability of each student with the course material, a minimum amount of work must be submitted to provide enough evidence of proficiency. To this end, students must submit the following assessments in order to be considered for a passing grade in the course: **the video project, the term test, and part 3 of the final project**. As these are summative assessments, if a student fails to submit one or more of these assessments (even if all other assessments have been completed), it will not be possible to gauge the student's proficiency with the material and will therefore not be able to pass the course.

EVALUATION BREAKDOWN

Quercus Discussion Participation: Group discussions will be conducted through the use of the Quercus discussion board. Every three weeks, a discussion topic will be posted based on content presented in the last and upcoming weeks of lectures (**see schedule at end of document for exact deadlines**). These topics will require students to discuss various applications of the course material and to think about how and why certain methods may be appropriate or not. Understanding the limitations of the statistical tools you use is what differentiates a good statistician from a great one! All students are encouraged to participate in these discussions for their participation grade.

- Topics will be open-ended (there is no one right answer - just join and engage with the discussion) and TAs and instructors will also be involved in these discussions.
- Each discussion will remain open for contribution for three weeks so it's best not to wait until the last minute to contribute.
- A rubric will be posted explaining how this will be graded.
- The first discussion will open January 14 and will be due on February 3 (see schedule for remaining deadlines).

In-class Group Labs: There will be 5 synchronous class periods during which a small group activity will take place. The activities will focus on getting hands-on practice applying the methods using R and writing up results. You and your group members will work together to perform a small data analysis to answer a question.

- Each lab will need to be turned in on Crowdmark by Thursday at 11:59PM EST the week of the lab (see schedule for exact dates) to receive completion credit. The additional time is in case of tight class schedules or if groups wish to work a little more on the lab.
- Students will need to ensure that the names of all group members who were present during the class are listed on the lab - these will help us ensure everyone gets credit for the work.
- Only 3 out of 5 labs need to be submitted to receive the full 5% in the final grade calculation (although it is encouraged to attempt all labs), however the best of both grading schemes will still be used for grade calculations so there is no penalty for not attending lecture/submitting the labs.

Weekly Quizzes: These quizzes will be available once each weekly module opens, and students can take the quiz at any time up until the Tuesday 11:59PM ET deadline. The quizzes will have a 1-hour time limit, although they should not take this long to complete. They will be multiple choice in nature and focus on the material covered in that week's module. Therefore you should watch the asynchronous module materials prior to completing the quiz. Only the **best 8 out of 10 quiz marks** will be used to calculate a student's final quiz grade. As such, no accommodations will be made for a missed quiz.

Reproducible Writing Exercise: This exercise is to highlight the importance of writing in science, specifically in a way that another independent researcher could reproduce what you have done based solely on a summary of your process. It also provides an opportunity for students to experience the scientific review and editing process. It will take place in three parts:

- Part 1 - Draft/Create: Students will submit a draft summary of a data analysis process that they applied to a dataset, for completion points.
- Part 2 - Peer Feedback/Assess: Students will have their draft reviewed by another student (peer) who will attempt to replicate their analysis. The reviewer student will provide comments on what is good and what could be improved with the draft. This will be graded for completion only.
- Part 3 - Final Draft/Revise: Students will revise their original draft, taking into account the feedback provided by their peer reviewer and submit their final product for grades. They will also rate the feedback provided to them by their reviewer based on helpfulness.

Term Test: The term test will be conducted online during the scheduled synchronous class time (see top of page 1). The test will be 1 hour and 30 minutes long, with an additional 20 minutes available to upload your solutions to Crowdmark. A link with the test questions will be emailed to students at the start time of the test, and all submissions must be received before the deadline to be graded. More details on submission will be communicated closer to the test date. The term test will take place during the scheduled synchronous lecture times on **February 16 from 5-7PM ET**. As per the timetable delivery instructions, students must be available during this time. The test will cover material from Modules 0-4.

Video Project: The purpose of the video project is to develop your data analysis skills which will be useful for the final project and future courses, in addition to your communication and statistical presentation skills. The video project 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 project will be:

- In groups of up to 2 students, use the methods taught in lecture to perform a data analysis and then present your results to a general audience.
- 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 need to display the results of your project in a logical way using slides or some other visual aid (e.g. PowerPoint, or other) and record you and your partner discussing these results, with a focus on why you chose to do certain things and interpretation of your results for a general (non-statistical) audience.
- Presentations should be submitted on time (i.e. by the deadline). Late submissions will receive a 10% penalty for each day that the project is late, up to 72 hours at which point the project will not be accepted. There will be a 1-hour grace period before the late penalty is applied, to account for unanticipated technical issues.
- In general, extensions will not be given unless you are experiencing a serious medical or personal emergency. If this is the case, you will be asked to complete a form available on Quercus to be submitted as early as possible (ideally before the deadline, but no later than 3 days after the deadline) to request an extension.
- **There is no make-up video project.** A missed video project will be given a grade of 0.

Final Project: The final project will be due during the final assessment period (date to be confirmed as soon as possible) and will consist of a data analysis on a novel dataset of your choice. Students will be required to demonstrate their understanding of the methods taught in lecture by developing a reasonable regression model that addresses a valid research question 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 is a scaffolded assessment involving 3 parts:

- Part 1- Research question and dataset selection: Students must find a dataset available online and define a research question that can be answered with this dataset using linear regression. Students will need to explain why their research question is important and how linear regression may be used to answer it. A short exploratory data analysis of the chosen dataset will also be required.
- Part 2 - Analysis Plan Flowchart: Students will be asked to put together a flowchart outlining the steps that they plan to take in their data analysis for the final project on their chosen dataset. This will help in developing a consistent analysis flow and make writing the final report easier.
- Part 3 - Final Project Report: Students will put together a scientific report that outlines the relevance of their proposed research question, the process of their analysis, the results of the performed data analysis, and a discussion of the meaning of the results as well as limitations of the analysis with respect to the statistical tools used/decisions made or the data used.

The final project will be done individually, and must be typed and submitted by the deadline. More detailed instructions will be provided at a later date. **In order to pass the course, you must submit part 3 of the final project.**

LATE ASSESSMENT POLICY

Assessment	Late Policy
Quizzes, Labs, Discussion, Term Test, Reproducible Exercise	no late submission accepted
Video Project	1-hour grace period, then 10% per each day late, up to maximum of 3 days late
Final Project Part 1 and 2	1-hour grace period, then 5% per each day late, up to maximum of 3 days late

Final Project Part 3	1-hour grace period, no late submissions accepted after that
----------------------	--

MISSED ASSESSMENT POLICY

If you experience a prolonged absence due to illness or emergency that prevents you from completing a number of assessments, please contact your registrar as soon as possible.

Extensions, if applicable to the assessment, will ONLY be considered for severe illness or emergencies, and requests citing e.g. busy schedules or many deadlines will not be considered.

Missed Discussion Board Participation: Participation is open for three weeks at a time and does not require a large time commitment to receive full marks. Therefore, no accommodations will be made for missed participation marks.

Missed In-class Labs: Since only the best 3 out of 5 labs count towards your lab grade and the labs can have a weight of 0% of the final grade under Scheme 2, there will be no accommodations for missed labs.

Missed Weekly Quiz: Students may miss up to 2 weekly quizzes in the term. These will be accommodated by having only 8 out of the 10 quizzes count towards the Quiz component of the final grade. No accommodations will be provided for any additional missed quizzes.

Missed Writing Exercise: Due to the scaffolded nature of this exercise, there will be no extensions on Parts 1 or 2 of this exercise. However, if a student is experiencing serious illness or personal emergency, an extension may be granted for Part 3. Submit a request using a form available on Quercus to be submitted as early as possible (ideally before the deadline, but no later than 3 days after the deadline) to request an extension.

Missed Video Project: There are no accommodations for a missed video project. However extensions may be granted to students experiencing serious personal illness or emergency at the discretion of the instructor. In this case, please submit a form available on Quercus to be submitted as early as possible (ideally before the deadline, but no later than 3 days after the deadline) to request an extension. The video project must be submitted as part of the minimum work requirement.

Missed Term Test: If a student is experiencing a serious personal illness or emergency on the date of the test, the student **must declare their absence on ACORN and notify the teaching team using a form available on Quercus no later than one week after the date of the test.** A make-up test may then be scheduled at a date and time determined by the instructor. **The format of the make-up is at the discretion of the instructor and could be multiple choice or an oral exam.** A few notes on missed term tests:

- To meet the minimum work requirement for this course, you must write the make-up if you missed the term test for a valid and documented medical reason, otherwise a meaningful grade cannot be calculated for you to pass the course.
- Since the term test is online and all students receive the test link, if you write the test or look at the test questions, you forfeit your eligibility to write the make-up test. Therefore you must make your decision about whether you are well enough to write the test before the test has begun.

Missed Final Project: The final project (part 3) must be completed in order to meet the minimum work requirement to pass the course, so no accommodations will be made for missing the final assessment.

Students will be given ample time to complete the assessment and extensions in general will not be granted.

REGRADE REQUESTS

Regrade requests will be accepted for an assessment worth 5% or higher (i.e. not the participation or weekly quizzes). Regrade requests must provide a justification for where there exists a grading error and/or how the work meets the grading rubric. These justifications must further be backed up with concrete references to the course material. All regrade requests will be accepted through a form available on the Quercus course page and will be accepted no later than one week after the grade for that assessment is released. **No regrade requests will be accepted by email or after the 1 week deadline.** The instructor further reserves the right to re-evaluate the assessment in its entirety (i.e. grades can go up, down, or remain unchanged).

INTELLECTUAL PROPERTY

Course materials provided on Quercus, such as lecture slides, assessments, videos and solutions are the intellectual property of your instructor and are for the use of students currently enrolled in this course only. Synchronous sessions will be recorded and be made available to other students enrolled in the course. **Providing course materials to any person or company outside of the course is unauthorized use and violates copyright.**

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:

- Being dishonest when reporting an illness or personal emergency to get an extension or accommodation is an academic offence.
- You may consult class notes/lecture slides during assessments, 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 unless otherwise specified.
- Paying anyone else to complete your assessments for you is academic misconduct.
- Completing assessments for another student is academic misconduct.
- Sharing your answers/work/code with others is academic misconduct.
- Using sources external to the course (anything not on Quercus) on an assessment 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, etc. 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 don't 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>.

TENTATIVE SCHEDULE OF TOPICS

Below is a tentative schedule of topics to be covered in class. The schedule is subject to change and modification.

Module (Dates)	Content
0 (Jan. 7-13)	Introduction: syllabus, review of mathematical/statistical and matrix concepts needed, introduction to R/RStudio/JupyterHub
1 (Jan. 14-20)	Good Data and Reporting Practices: good data exploration, good reporting practices, good communication practices
2 (Jan. 21-27)	The Regression Model: model specification, least squares estimation, interpretation
3 (Jan. 28-Feb.3)	Assumptions and Properties of Residuals and Estimators: LS assumptions, sampling distributions of parameters, impact of model violations
4 (Feb. 4-10)	Confidence and Tests in Regression: CIs and tests on the intercept, slope, mean response, and prediction intervals for an actual response
Feb. 11-17	No new material - term test
Feb. 21-25	READING BREAK
5 (Feb. 25-Mar. 3)	Decomposing the Variance: Sum of squares decomposition, coefficient of determination, ANOVA and ANCOVA F tests, Partial F test
6 (Mar. 4-10)	Identifying and Mitigating Violated Assumptions: residual plot diagnostics, transformations for non-constant variance and non-linearity
7 (Mar. 11-17)	Problematic Observations: Outliers, Leverage Points, Influential Points, Detection and Impact
8 (Mar. 18-24)	Related Covariates: Multicollinearity and VIF, interaction terms, basics of model selection using coefficients of determination
9 (Mar. 25-31)	Variable Selection Techniques: Numerical selection criteria, stepwise selection procedures, cautions and the use of context
10 (Apr. 1-7)	Model Validation and Wrap-up: How to validate your models, MLR data analysis process overview and report guidelines
Apr. 11-29	Final assessment period - Final project due tentatively on Apr. 20

CALENDAR OF DATES AND DEADLINES

For a complete list of due dates and synchronous class times, see the attached calendar. It is recommended that you save/print the calendar and/or copy the dates to your personal calendar to make it easier to stay on track with the course.

JAN2022

SUN

MON

TUE

WED

THU

FRI

SAT

01

02

03

04

05

06

Module 0
released

07

08

09

10

11

Lecture (5-7PM
ET)

12

13

Module 1
released +
Discussion 1
opens

14

15

16

17

Quiz 1 due by
11:59PM ET

18

Lecture (5-7PM
ET)

19

Lab 1 due by
11:59PM ET

20

Module 2
released

21

22

23

Last day to
enroll

24

Quiz 2 due by
11:59PM ET

25

Lecture (5-7PM
ET)

26

Lab 2 due by
11:59PM ET
Rep. Ex. Create
due by 11:59PM
ET

27

Module 3
released

28

29

30

Rep. Ex. Assess
due by 11:59PM
ET

31

FEB 2022

SUN

MON

TUE

WED

THU

FRI

SAT

01

02

03

04

05

Quiz 3 due by
11:59PM ET

Lecture (5-7PM
ET)

Discussion 1
due 11:59PM
ET

Module released
+ Discussion 2
opens
Rep. Ex. Revise
due 11:59PM

06

07

08

09

10

11

12

Quiz 4 due by
11:59PM ET

Lecture (5-7PM
ET)

Lab 3 due by
11:59PM ET

13

14

15

16

17

18

19

Term test (5-
7PM ET)

20

21

22

23

24

25

26

Final Project Part
1 due by
11:59PM ET

————READING BREAK————

Discussion 2 due
11:59PM ET

Module 5
released +
Discussion 3
opens

27

28

MAR2022

SUN

MON

TUE

WED

THU

FRI

SAT

01

02

03

04

05

Quiz 5 due by
11:59PM ET

Lecture (5-7PM
ET)

Module 6
released

06

07

08

09

10

11

12

Quiz 6 due by
11:59PM ET

Lecture (5-7PM
ET)

Lab 4 due by
11:59PM ET

Module 7
released

13

14

15

16

17

18

19

Drop deadline

Quiz 7 due by
11:59PM ET

Lecture (5-7PM
ET)

Discussion 3
due 11:59PM
ET

Module 8 released
+ Discussion 4
opens
Video project due
11:59PM ET

20

21

22

23

24

25

26

Quiz 8 due by
11:59PM ET

Lecture (5-7PM
ET)

Lab 5 due by
11:59PM ET

Module 9
released

27

28

29

30

31

Quiz 9 due by
11:59PM ET

Lecture (5-7PM
ET)

APR 2022

SUN

MON

TUE

WED

THU

FRI

SAT

01

02

Module 10
released
Final Project Part
2 due by 11:59PM
ET

03

04

05

06

07

08

09

Quiz 10 due by
11:59PM ET

Lecture (5-7PM
ET)

Discussion 4
due 11:59PM
ET

Last day of class

10

11

12

13

14

15

16

17

18

19

20

21

22

23

Final project part
3 due 11:59PM
ET

24

25

26

27

28

29

30

*End of final
assessment
period*