Statistical Methods for Data Analytics

MSCI 718, Winter 2022

Instructor: Dr. Jangho Yang
Email: j634yang@uwaterloo.ca
Website: https://janghoyang.com
Class: Tue & Thur 2:30 pm - 3:50 pm, E2 1736 (online until Feb 7th)
Office Hour: Wednesday, 10 am– 12:30 pm, and by appointment, CPH 3629 (online only until Feb 7th)

Teaching Assistants: Yekta Amirkhalili & Muhammad Saadi Aziz
Emails: yekta.amirkhalili@uwaterloo.ca & m29aziz@uwaterloo.ca

Course Description

The objective of this course is to develop skills with a range of procedures and programs for multivariate data analysis. The focus will be on practical issues such as selecting the appropriate analysis, preparing data for analysis, menu-driven and syntax programming, interpreting output, and presenting results of a complex nature. The course aims for an intuitive understanding of quantitative methods based on examples and the actual implementation of methods using statistical programming, and therefore the use of matrix algebra is limited unless necessary. Prerequisites are introductory statistics courses such as MSCI 609.

Course Objectives

1. Improve proficiency in R programming language for statistical modeling and computation.
2. Select and apply a variety of statistical tools to answer quantitative research questions and formalize certainty in those answers
3. Design experiments and statistical models to represent quantitative research questions
4. Analyze and communicate the findings of statistical tools

Main Readings

Main textbook

- [Y] J. Yang, Lecture Notes: Quantitative Data Analysis (Spring 2021). Get a pdf here.
- [G] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021) , *An Introduction to Statistical Learning: with Applications in R* (2nd Edition), New York: springer. Get a free pdf here.

Supplementary textbooks

Overview of key concepts in statistics:

- Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2006), *Probability and Statistics for Engineers and Scientists* (9th Edition), Pearson.

- Casella, G., & Berger, R. L. (2002), *Statistical Inference* (2nd Edition), Duxbury Press.

Overview of modern statistical methods, including those widely used in machine learning:

- Hastie, T., Tibshirani, R., & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media

Mathematical foundations of statistics:

- Hogg, R. V., McKean, J., & Craig, A. T. (2005), *Introduction to Mathematical Statistics* (7th Edition), Pearson Education.

Probability concepts from the information-theoretical point of view:

- Cover, T. and Thomas, J.(1999). *Elements of Information Theory*, John Wiley & Sons

Introductory & Intermediate Bayesian statistics:

- McElreath, R. (2020), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2nd Edition), CRC press.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013), *Bayesian Data Analysis* (3rd Edition). CRC press

Software and Programming

The course will involve empirical analysis and require the use of statistical software. R is the main statistical software for this course but other programming software will be allowed if necessary. An Integrated Development Environment (IDE) will be useful for handling code, debugging, and sourcing control. For R users, Rstudio is recommended. R is an open-source programming language and is available for free download. To download R, go to https://www.r-project.org/.

Course Delivery and Communication

Lectures and Live Session

- The lectures will be live streamed on Teams on Tuesdays and Thursdays. A recording will be distributed until Feb 7th.
- A weekly live coding session will be run by TAs on Thursdays from 3:05-3:50 pm. Depending on the progress of students, the coding session will be held bi-weekly after the midterm. The live-coding session will be live-streamed and recorded throughout the term.
- The course will go back to fully in-person after Feb 7th unless otherwise announced.

Assignment/project submission and in-class test

- Assignments and the final project need to be submitted through Learn.
- Midterm exam is held onilne.

Announcements

- All the course announcements will be made on Learn.

Online discussions and questions

- There are two forums for discussions and questions created on Teams: Weekly Topic Discussions & General Course Discussion. Use the Weekly Topic Discussions forum to ask questions related to course materials each week. Use the General Course Discussion Forum to ask clarifying questions related to course logistics, e.g. exam schedule, assignment due date. The instructor team will regularly check and respond to all questions as quickly as possible.

Email Policy

- All course related inquiries should be posted on Teams. Course related questions either about course materials and course logistics sent to the instructor team via emails will be returned to the student with the statement "Please post this question on Teams." This policy is motivated by the non-rivalrous and non-excludable nature of the intellectual communications between students and the instructor team. All students enrolling in this course have the equal right to participate in collective learning through questions and answers.
- E-mail is an official means for communication only when i) students have personal issues to be discussed with the professor, e.g. accommodations due to extenuating circumstances or assignment/exam grading appeal, and ii) students have confidential feedback to the instructor team.

Course Requirements and Marking Scheme

Assignments	40%	(20% each)
Midterm exam	30%	
Final project	30%	

Notice:

1. The two assignments include programming exercises to be solved in R and short essay questions about statistical concepts. Assignments need to be submitted through Learn.
2. The midterm exam is held online.

3. There will be 3 final project topics. Students need to choose one topic and submit a final report. More details will be discussed in class.
4. Lateness policy for assignment: Each assignment is graded 0-100 points. A late assignment gets 10 points deduction per 12 hour, rounded up. Late assignments are only accepted for 48 hours after the deadline. Late submission due to extenuating circumstances is exempted from the point deduction. Extenuating circumstances include an extended illness requiring hospitalization or visit to a physician with documentation and a family emergency, e.g. serious illness (with written explanation). Students need to submit a University of Waterloo Verification of Illness Form. Please refer to Accommodation due to Illness Policy for more information.
5. Late submission for the final project is not accepted unless there are extenuating circumstances.
6. No plagiarism is tolerated in any circumstances. See Academic Integrity below for more information.
7. If you are unable to attend a session or meet a deliverable deadline, please let our teaching team know immediately. If you are facing challenges that are affecting more than one course, please contact your Associate Chair or Director of your program. They will review your case and coordinate a reasonable and fair plan in consultation with appropriate others (for example: instructors, Department Undergraduate Studies Committee, Chair, AccessAbility Services, Engineering Counselling services, Registrar's Office).

Fair Contingencies for Emergency Remote Teaching.

We are facing unusual and challenging times. The course outline presents the instructor's intentions for course assessments, their weights, and due dates in Winter 2022. As best as possible, we will keep to the specified assessments, weights, and dates. To provide contingency for unforeseen circumstances, the instructor reserves the right to modify course topics and/or assessments and/or weight and/or deadlines with due and fair notice to students. In the event of such challenges, the instructor will work with the Department/Faculty to find reasonable and fair solutions that respect rights and workloads of students, staff, and faculty.

Covid-19 Emergency Remote Teaching-Learning

In the case of another COVID outbreak, we wil go back to online format. Live Sessions will be held on MS Teams during scheduled course times and the sessions will be recorded.

Important Dates

Class begins	Jan 05
1st Assignment due	Feb 3
Reading Week	Feb 19 - Feb 27
Midterm Exam	Mar 03
2nd Assignment due	Mar 21
Class Ends	Apr 5
Final Project Submission	Apr 8

Topics and Readings

Topic 1: Introduction and Probabilistic Thinking with R

Week 1-3 (Jan 6, 11, 13, & 18)	Y. Ch.1 G. Ch.1-2	• Probability, joint & conditional probability • Bayes' theorem • Probability distribution and RNG • Examples: Simple urn, Polya urn, Monty hall problem, and stick breaking

Topic 2: Model fitting and comparison

Week 3 & 4 (Jan 25 & Feb 1)	Y. Ch.2 G. Ch.2.1, 2.2, 5	• Fitting exercise: Normal, exponential, and power law distributions • Maximum Likelihood Estimation (MLE) • Bias-variance tradeoff • In-sample vs. out-of-sample prediction

Topic 3: Simple Linear Regression

Week 5 (Feb 3 & 8)	Y. Ch.4 G. Ch.3.1	• Basic linear model specification with one predictor and one response variable • Example: α and β in the stock market • Log linear and interpretation • Normality assumption of error and its justification/falsification • Post-estimation evaluation: residual analysis and goodness of fit

Topic 4: Multiple Linear Regression

Week 6 & 7 (Feb 10, 15, & 17)	Y. Ch.5 G. Ch.3.2	• Simpson's paradox • Example 1: College admission rates • Hidden causation • Example 2: Mother's age and child's test score • Interaction • Example 3: Training effect on wage

Midterm Exam

Week 9
(March 3)

Topic 5: Generalized Linear Model and Classification

Week 10 (Mar 8 & 10)	Y. Ch.6 G. Ch.4	• Link function • Logistic regression • Example 1: Equality of opportunity and innovation • Poisson regression • Example 2: Stop and frisk

Topic 6: Multilevel Model and Regularization

Week 11 & 12 (Mar 15, 22 &24)	Y. Ch.7 G. Ch.6	• James–Stein estimator • Example 1: Baseball prediction • Varying intercept and slope coefficients • Example 2: Cafe wait time • Example 3: Corporate investment rate

Project discussion

Week 11 (Mar 17)		• Final project guidelines

Topic 7: Mixture models: Examples

Week 13 (Mar 29)		• Chinese restaurant and Indian restaurant problem • Gaussian processes • Dirichlet process

Final Project

Week 14 (Apr 8)		

Academic Integrity and Students with Disabilities

Academic Integrity

In order to maintain a culture of academic integrity, members of the University of Waterloo community are expected to promote honesty, trust, fairness, respect and responsibility. Check the Office of Academic Integrity's website for more information.

All members of the UW community are expected to hold to the highest standard of academic integrity in their studies, teaching, and research. This site explains why academic integrity is important and how students can avoid academic misconduct. It also identifies resources available on campus for students and faculty to help achieve academic integrity in — and out — of the classroom.

Intellectual Property

Students should be aware that this course contains the intellectual property of their instructor, TA, and/or the University of Waterloo. Intellectual property includes items such as:

- Lecture content, spoken and written (and any audio/video recording thereof)
- Lecture handouts, presentations, and other materials prepared for the course (e.g., PowerPoint slides)
- Questions or solution sets from various types of assessments (e.g., assignments, quizzes, tests, final exams)
- Work protected by copyright (e.g., any work authored by the instructor or TA or used by the instructor or TA with permission of the copyright owner).

Course materials and the intellectual property contained therein, are used to enhance a student's educational experience. However, sharing this Intellectual property without the intellectual property owner's permission is a violation of intellectual property rights. For this reason, it is necessary to ask the instructor, TA and/or the University of Waterloo for permission before uploading and sharing the intellectual property of others online (e.g., to an online repository).

Permission from an instructor, TA or the University is also necessary before sharing the intellectual property of others from completed courses with students taking the same/similar courses in subsequent terms/years. In many cases, instructors might be happy to allow distribution of certain materials. However, doing so without expressed permission is considered a violation of intellectual property rights.

Grievance

A student who believes that a decision affecting some aspect of his/her university life has been unfair or unreasonable may have grounds for initiating a grievance. Read Policy 70 — Student Petitions and Grievances, Section 4. When in doubt please be certain to contact the department's administrative assistant who will provide further assistance.

Discipline

A student is expected to know what constitutes academic integrity, to avoid committing academic offenses, and to take responsibility for his/her actions. A student who is unsure whether an action constitutes an offense, or who needs help in learning how to avoid offenses (e.g., plagiarism, cheating) or about "rules" for group work/collaboration should seek guidance from the course professor, academic advisor, or the Undergraduate Associate Dean. For information on categories of offenses and types of penalties, students should refer to Policy 71 — Student Discipline. For typical penalties, check

Guidelines for the Assessment of Penalties.

Avoiding Academic Offenses

Most students are unaware of the line between acceptable and unacceptable academic behaviour, especially when discussing assignments with classmates and using the work of other students. For information on commonly misunderstood academic offenses and how to avoid them, students should refer to the Faculty of Mathematics Cheating and Student Academic Discipline Policy.

Appeals

A decision made or a penalty imposed under Policy 70, Student Petitions and Grievances (other than a petition) or Policy 71, Student Discipline may be appealed if there is a ground. A student who believes he/she has a ground for an appeal should refer to Policy 72 — Student Appeals.

Note for students with disabilities

The AccessAbility office is located in Needles Hall, Room 1401, collaborates with all academic departments to arrange appropriate accommodations for students with disabilities without compromising the academic integrity of the curriculum. If you require academic accommodations to lessen the impact of your disability, please register with AccessAbility Services at the beginning of each academic term.

Turnitin.com

Text matching software (Turnitin) may be used to screen assignments in this course. Turnitin is used to verify that all materials and sources in assignments are documented. Students' submissions are stored on a U.S. server, therefore students must be given an alternative (e.g., scaffolded assignment or annotated bibliography), if they are concerned about their privacy and/or security. Students will be given due notice, in the first week of the term and/or at the time assignment details are provided, about arrangements and alternatives for the use of Turnitin in this course.

It is the responsibility of the student to notify the instructor if they, in the first week of term or at the time assignment details are provided, wish to submit alternate assignment.

Academic Accommodations

Fair Contingencies for Emergency Remote Teaching

To provide contingency for unforeseen circumstances, the instructor reserves the right to modify course topics and/or assessments and/or weight and/or deadlines with due notice to students. In the event of further challenges, the instructor will work with the Department/Faculty to find reasonable and fair solutions that respect rights and workloads of students, staff, and faculty.

Online Academic Integrity for Individual Assessments

For all graded course assessments, students are expected to work individually and submit their own original work. Under Policy 71, the instructor may have follow-up conversations with individual students to ensure that the work submitted was completed on their own. Any follow up will be conducted remotely (e.g., MS Teams, Skype, phone), as the University of Waterloo has suspended all in-person meetings until further notice. Any permissions for collaboration on assessments (e.g., team project) must be provided by the instructor in writing.

Compassionate Consideration

If you are facing challenges that are affecting more than one course, please contact your Associate Chair or Director of your program. They will review your case and coordinate a reasonable and fair plan in consultation with appropriate others (for example: Instructors, Department Undergraduate Studies Committee, Chair, AccessAbility Services, Engineering Counselling services, Registrar's Office).

Wellness Support and Contact Information

We all need a support system. We encourage you to seek out mental health supports when they are needed. Please reach out to Campus Wellness and Counselling Services. We understand that these circumstances can be troubling, and you may need to speak with someone for emotional support. Good2Talk is a post-secondary student helpline based in Ontario, Canada that is available to all students including outside Ontario. MATES is a one-to-one student peer support program offered by the Waterloo Undergraduate Student Association in consultation with Campus Wellness. MATES provides support to students who are hoping to build social skills, or are experiencing personal or academic concerns or low-level mental health and wellness difficulties.

Appendix

All engineering programs are reviewed by the Canadian Engineering Accreditation Board (CEAB). One of the required accreditation criteria is that institutions ensure students have sufficient knowledge and proficiency with respect to the 12 Graduate Attributes (GAs) listed below. These attributes are mapped to the learning objectives in each course for assessment, as shown in the brackets. This allows the program to both comply with CEAB requirements and continuously improve

#	Acronym	Attribute Name	Attribute Definition
1	KB	Knowledge Base	Demonstrated competence in university level mathematics, natural sciences, engineering fundamentals, and specialized engineering knowledge appropriate to the program.
2	PA	Problem analysis	An ability to use appropriate knowledge and skills to identify, formulate, analyze, and solve complex engineering problems in order to reach substantiated conclusions.
3	Inv	Investigation	An ability to conduct investigations of complex problems by methods that include appropriate experiments, analysis and interpretation of data, and synthesis of information in order to reach valid conclusions.
4	Des	Design	An ability to design solutions for complex, open-ended engineering problems and to design systems, components or processes that meet specified needs with appropriate attention to health and safety risks, applicable standards, and economic, environmental, cultural and societal considerations.
5	Tools	Use of Engineering Tools	An ability to create, select, apply, adapt, and extend appropriate techniques, resources, and modern engineering tools to a range of engineering activities, from simple to complex, with an understanding of the associated limitations.
6	Team	Individual and team work	An ability to work effectively as a member and leader in teams, preferably in a multi-disciplinary setting.
7	Comm	Communication skills	An ability to communicate complex engineering concepts within the profession and with society at large. Such ability includes reading, writing, speaking and listening, and the ability to comprehend and write effective reports and design documentation, and to give and effectively respond to clear instructions.
8	Prof	Professionalism	An understanding of the roles and responsibilities of the professional engineer in society, especially the primary role of protection of the public and the public interest.
9	Impact	Impact of engineering	An ability to analyze social and environmental aspects of engineering activities. Such ability includes an understanding of the interactions that engineering has with the economic, social, health, safety, legal, and cultural aspects of society, the uncertainties in the prediction of such interactions; and the concepts of sustainable design and development and environmental stewardship.
10	Ethics	Ethics and equity	An ability to apply professional ethics, accountability, and equity.
11	Econ	Economics and project management	An ability to appropriately incorporate economics and business practices including project, risk, and change management into the practice of engineering and to understand their limitations.
12	LL	Life-long learning	An ability to identify and to address their own educational needs in a changing world in ways sufficient to maintain their competence and to allow them to contribute to the advancement of knowledge.