*We recognize and acknowledge that McMaster University meets and learns on the traditional territories of the Mississauga and Haudenosaunee nations, and within the lands protected by the “Dish With One Spoon” wampum, an agreement amongst all allied Nations to peaceably share and care for the resources around the Great Lakes.*

# CSE/STATS 790 – Statistical learning

# 2023 Winter Term

**Instructor:**  Dr. Ben Bolker **Office hours:** TBD/by appointment, HH 314 **E-mail:** [bolker@mcmaster.ca](mailto:formann@mcmaster.ca)

**Lectures:** M 3:30-5:30, W 9:30-10:30

## Course Description

* Learn new data science methods, focusing on regression (i.e., supervised learning for predictions of continuous outcomes).
* Study the bias-variance trade-oﬀ and eﬀective degrees of freedom in statistical prediction; how different methods optimize the bias-variance tradeoff; and how optimization methods are chosen for their computational efficiency and scalability.
* Learn how these methods work and how to implement most of them (i.e., from scratch).
* Practice implementing cross-validation and using it for model selection.
* Gain expertise with programming in R. Develop basic skills in Julia.

This course is in part a continuation of STAT/CSE 780, in that we continue to introduce more data science methods, though 780 is not a formal prerequisite. We will discuss methods in greater depth, to gain more understanding of how the method works and is implemented.

**Prerequisite(s):** While there are no formal prerequisites, to succeed in the course you will need to be comfortable with (1) multivariable calculus; (2) linear algebra (focusing on numerical computation, e.g. at the level of [MATH 3NA3](https://mcmasteru365.sharepoint.com/:w:/s/sci-course-outlines/ETS28ygy6rZBs3-oLnuLkRYBMthhG_x681iQsnGSbPjHMw?e=d3VrR8)); (3) basics of linear regression (e.g. at the level of [STATS 3A03](https://mcmasteru365.sharepoint.com/:w:/s/sci-course-outlines/EX5q-QIB-lRGvCleB8YStCIB1pfpenDKNb4Uc-53WEGBrg?e=ScAtu8)); (4) Some programming language; R is recommended, but knowledge of Python/Matlab/Java should be transferable; (5) basic knowledge of data science methods (e.g. lasso/ridge regression, random forests, boosting), at the level of STATS/CSE 780 (recommended).

**Antirequisite(s):** None.

## Course and Learning Objectives

### Learning Objectives

**Upon completion of this course, the student will be able to:**

1. Implement a range of estimation/prediction algorithms in R, including, but not limited to: k-nearest-neighbors smoothing, kernel smoothing, general cubic splines, smoothing splines, ridge regression, LASSO regression, Generalized additive models (GAMs), and Multivariate adaptive regression splines (MARS).
2. Explain the bias-variance trade-off and its connection to overfitting.
3. Explain a variety of data science methods for avoiding overfitting.

## Materials & Fees

### Required Materials / Resources

**Free required E-texts:**

* T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, 2nd ed. <https://web.stanford.edu/~hastie/Papers/ESLII.pdf>
* Cosma Shalizi. *Advanced Data Analysis from an Elementary Point of View.* <https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/>

**Other useful references:**

* [*R for Data Science*](https://r4ds.had.co.nz/)by Garrett Grolemund and Hadley Wickham.
* [*R Markdown Cookbook*](https://bookdown.org/yihui/rmarkdown-cookbook/)by Yihui Xie, Christophe Dervieux, Emily Riederer.

### Websites

* Course website on Avenue2Learn
* Assignments collected via Crowdmark.com
* A course Discord server, for discussion

### Computer

* A computer that meets performance requirements [found here](https://cto.mcmaster.ca/technology-resources-for-mcmaster-students/" \l "tab-content-device-recommendations).
* An internet connection, needed to download data and upload completed assignments.

If you think that you will not be able to meet these requirements, please contact [uts@mcmaster.ca](mailto:uts@mcmaster.ca) as soon as you can. Please visit the [Technology Resources for Students page](https://cto.mcmaster.ca/technology-resources-for-mcmaster-students/" \l "tab-content-device-recommendations) for detailed requirements. If you use assistive technology or believe that our platforms might be a barrier to participating, please contact [Student Accessibility Services](https://sas.mcmaster.ca/), [sas@mcmaster.ca](mailto:sas@mcmaster.ca), for support.

### Software

This course uses various free software (R, Julia, VS Code, Quarto, TeX). Instructions for installation will be posted at <https://github.com/bbolker/stat790/tree/main/software> ; you are strongly encouraged to work through these instructions and install software before the first class meeting.

## Course Overview and Assessment

### Topics

Tentative content/order of topics, subject to change.

|  |  |  |  |
| --- | --- | --- | --- |
| Week | Start date | Topic | Assignments due  (Fridays of corresponding week) |
| 1 | 9 Jan | Overview: big picture, software, bias-variance tradeoff |  |
| 2 | 16 Jan | Regression methods (OLS/lasso/ridge) | HW 1 |
| 3 | 23 Jan | Regression with non-Gaussian outcomes |  |
| 4 | 30 Jan | Basis expansion and regularization: GAMs | HW 2 |
| 5 | 6 Feb | Model assessment and selection |  |
| 6 | 13 Feb | Assessing uncertainty | HW 3 |
|  | 20 Feb | Midterm recess |  |
| 7 | 27 Feb | Tree-based methods (CART/MARS/etc.) |  |
| 8 | 6 Mar | Random forests, boosting | HW 4 |
| 9 | 13 Mar | Optimization and scalability | Project proposal |
| 10 | 20 Mar | Latent structures: factor, mixture, graphical models | HW 5 |
| 11 | 27 Mar | Data with dependence (temporal/spatial) |  |
| 12 | 3 Apr | Students’ choice |  |
| 13 | 10 Apr | Wrap up | Project draft |
|  | 28 Apr |  | Project final version |

## Presentation

This course will run fully in-person in accordance with university directives and strict health and safety guidelines. Students are generally expected to attend all lectures and tests in-person. However, students must be prepared to move to virtual learning should there be a change to health regulations and restrictions as issued by the Province or University.

## Evaluation

Assignments will be collected electronically via Crowdmark.com.

**Homework (75%)**: There will be n ≈ 5 homework assignments, each worth (75%/n) ≈ 15% of the total grade. These will typically involve introducing a data set and analyzing it with one or two methods from the lectures, then evaluating the success of these methods.

**Final project (25%)**: There will be a ﬁnal project due by 11:59 PM on Friday, April 28th. This

will be like a homework assignment but more open-ended and longer. The final submission will be preceded by a proposal submission and a first draft, which will be reviewed and returned with feedback.

**Regrades:** If you need more information on how an assignment was graded or wish to request a regrade, contact the instructor or TA via email.

**Collaboration and academic honesty:** The university’s formal statement on academic integrity is included below. Here are my own thoughts and guidelines speciﬁc to this course. You are encouraged to form study groups and collaborate on homework. However, you must each produce your own solutions. Do not copy solutions from the internet or a classmate. If you do so, in addition to violating the university’s policy on academic integrity, you will have compromised your own learning, and the grader must waste time evaluating work that does not reﬂect your progress. If you can’t complete a problem, just write a note explaining what you tried and why you think it didn’t work; an honest eﬀort will be rewarded with partial credit.

## Requests for Relief for Missed Academic Term Work

[McMaster Student Absence Form (MSAF):](https://secretariat.mcmaster.ca/university-policies-procedures-guidelines/msaf-mcmaster-student-absence-form/) In the event of an absence for medical or other reasons, students should review and follow the Academic Regulation in the Undergraduate Calendar “Requests for Relief for Missed Academic Term Work”.

## Academic Accommodation of Students with Disabilities

Students with disabilities who require academic accommodation must contact [Student Accessibility Services (SAS](https://sas.mcmaster.ca/)) at 905-525-9140 ext. 28652 or [sas@mcmaster.ca](mailto:sas@mcmaster.ca) to make arrangements with a Program Coordinator. For further information, consult McMaster University’s [*Academic Accommodation of Students with Disabilities*](https://secretariat.mcmaster.ca/app/uploads/Academic-Accommodations-Policy.pdf) policy.

## Academic Accommodation for Religious, Indigenous Or Spiritual Observances (Riso)

Students requiring academic accommodation based on religious, indigenous or spiritual observances should follow the procedures set out in the [RISO](https://secretariat.mcmaster.ca/app/uploads/2019/02/Academic-Accommodation-for-Religious-Indigenous-and-Spiritual-Observances-Policy-on.pdf) policy. Students should submit their request to their Faculty Office ***normally within 10 working days*** of the beginning of term in which they anticipate a need for accommodation or to the Registrar's Office prior to their examinations. Students should also contact their instructors as soon as possible to make alternative arrangements for classes, assignments, and tests.

## Courses with An On-Line Element

**Some courses may**use on-line elements (e.g. e-mail, Avenue to Learn (A2L), LearnLink, web pages, capa, Moodle, ThinkingCap, etc.). Students should be aware that, when they access the electronic components of a course using these elements, private information such as first and last names, user names for the McMaster e-mail accounts, and program affiliation may become apparent to all other students in the same course. The available information is dependent on the technology used. Continuation in a course that uses on-line elements will be deemed consent to this disclosure. If you have any questions or concerns about such disclosure, please discuss this with the course instructor.

## Online Proctoring

**Some courses may**use online proctoring software for tests and exams. This software may require students to turn on their video camera, present identification, monitor and record their computer activities, and/or lock/restrict their browser or other applications/software during tests or exams. This software may be required to be installed before the test/exam begins.

## Academic Integrity

You are expected to exhibit honesty and use ethical behaviour in all aspects of the learning process. Academic credentials you earn are rooted in principles of honesty and academic integrity.

**It is your responsibility to understand what constitutes academic dishonesty.**

Academic dishonesty is to knowingly act or fail to act in a way that results or could result in unearned academic credit or advantage. This behaviour can result in serious consequences, e.g. the grade of zero on an assignment, loss of credit with a notation on the transcript (notation reads: “Grade of F assigned for academic dishonesty”), and/or suspension or expulsion from the university. For information on the various types of academic dishonesty please refer to the [*Academic Integrity Policy*](https://secretariat.mcmaster.ca/app/uploads/Academic-Integrity-Policy-1-1.pdf)*,* located at <https://secretariat.mcmaster.ca/university-policies-procedures- guidelines/>

**The following illustrates only three forms of academic dishonesty:**

* plagiarism, e.g. the submission of work that is not one’s own or for which other credit has been obtained.
* improper collaboration in group work.
* copying or using unauthorized aids in tests and examinations.

## Authenticity / Plagiarism Detection

**Some courses may**use a web-based service (Turnitin.com) to reveal authenticity and ownership of student submitted work. For courses using such software, students will be expected to submit their work electronically either directly to Turnitin.com or via an online learning platform (e.g. A2L, etc.) using plagiarism detection (a service supported by Turnitin.com) so it can be checked for academic dishonesty.

Students who do not wish their work to be submitted through the plagiarism detection software must inform the Instructor before the assignment is due. No penalty will be assigned to a student who does not submit work to the plagiarism detection software. **All submitted work is subject to normal verification that standards of academic integrity have been upheld** (e.g., on-line search, other software, etc.). For more details about McMaster’s use of Turnitin.com please go to the [McMaster Office of Academic Integrity](https://www.mcmaster.ca/academicintegrity/)’s webpage.

## Conduct Expectations

As a McMaster student, you have the right to experience, and the responsibility to demonstrate, respectful and dignified interactions within all our living, learning and working communities. These expectations are described in the [*Code of Student Rights & Responsibilities* (the “Code”).](https://secretariat.mcmaster.ca/app/uploads/Code-of-Student-Rights-and-Responsibilities.pdf) All students share the responsibility of maintaining a positive environment for the academic and personal growth of all McMaster community members, **whether in person or online**.

It is essential that students be mindful of their interactions online, as the Code remains in effect in virtual learning environments. The Code applies to any interactions that adversely affect, disrupt, or interfere with reasonable participation in University activities. Student disruptions or behaviours that interfere with university functions on online platforms (e.g. use of Avenue 2 Learn, WebEx or Zoom for delivery), will be taken very seriously and will be investigated. Outcomes may include restriction or removal of the involved students’ access to these platforms.

## Copyright and Recording

Students are advised that lectures, demonstrations, performances, and any other course material provided by an instructor include copyright protected works. The Copyright Act and copyright law protect every original literary, dramatic, musical and artistic work, **including lectures** by University instructors.

The recording of lectures, tutorials, or other methods of instruction may occur during a course. Recording may be done by either the instructor for the purpose of authorized distribution, or by a student for the purpose of personal study. Students should be aware that their voice and/or image may be recorded by others during the class. Please speak with the instructor if this is a concern for you.

## Extreme Circumstances

The University reserves the right to change the dates and deadlines for any or all courses in extreme circumstances (e.g., severe weather, labour disruptions, etc.). Changes will be communicated through regular McMaster communication channels, such as McMaster Daily News, A2L and/or McMaster email.