Course Description

MATH 4939: Statistical Data Analysis using SAS and R – Winter 2021

Georges Monette

January 2021

Contents

Technical Requirements **1**

Goals of MATH 4939 **2**

Course work and grades **2**

Prerequisites **4**

References **4**

Getting Help **4**

Some reflections on teams **5**

Course policies **5**
Missed deadlines 5
Missed term test 5
Academic honesty 5

Bibliography **5**

(Updated: July 10 2021 11:40)

> Doubt is not a pleasant condition, but certainty is absurd. — *Voltaire*

Technical Requirements

There are technical requirements to be able to complete this course. You need a reliable high-speed internet connection to attend lectures via Zoom and to participate in quizzes. You need a webcam or a phone with camera to take part in oral tests and for the oral presentation connected with the project.
You need a scanner or a phone with a camera to scan and upload quizzes.

For more information, see:

- https://lthelp.yorku.ca/student-guide-to-moodle
- https://yorku.zoom.us/ (see the guides at the bottom)
- https://uit.yorku.ca/students-getting-started/ and Student Guide to eLearning

Goals of MATH 4939

MATH 4939 is a 4th year capstone course. You have already taken a number of statistics courses that focus on diverse methodologies that are applied to problems and data suited to the specific methodology of each course.

The fundamental purpose of MATH 4939 is to bring together all that you have previously learned in statistics to provide you with the skills to apply this knowledge effectively, creatively and correctly to the task of addressing real scientific or business questions with real data.

When you succesfully complete this course you should have improved your ability to:

- describe the factors that differentiate between observational and experimental data and know what questions to ask to identify the kind of data you are working with correctly,
- differentiate among different types of scientific and business questions: predictive, causal/explanatory and descriptive questions,
- identify the distinct roles of explanatory variables and relate them to their positions in causal graphs,
- adapt model building strategies and model evaluation strategies appropriately for the type of scientific or business question, the type of data being analyzed, and the roles of available variables,
- be able to express limitations in the interpretation of results due to possible omitted variables, incorrect model specification, incorrectly included variables that may bias model interpretation, incorrect data, etc.,
- know what questions to ask to identify the presence and nature of missing data, including data whose missingness may not be apparent in a data set,
- apply basic strategies to deal with missing data,
- identify common statistical fallacies you may have acquired that arise from generalizing principles that appear reasonable in the context of simple linear additive problems commonly used in introductory courses but that lead to consequential errors when applied to complex data analysed with the goal of addressing real scientific questions,
- learn techniques and computer coding to visualize data and to visualize statistical inferences and to apply them appropriately in the analysis and reporting of conclusions,
- learn the theory and application of hierarchical models that allow a synthesis of methodologies learned in previous courses so they can be applied in a common modelling strategy,
- communicate statistical analyses and findings in a manner that is appropriate for different audiences.

Course work and grades

- **Quizzes: 25%** Weekly quizzes on Wednesdays.
  - Late quizzes will be subject to a penalty that will be announced before the beginning of the quiz.
  - The lowest grade is dropped.
  - If you can't attend a quiz for medical or other reasons beyond your control, the weight of the quiz will be transferred to the final oral exam.
- **Mid-term oral: 10%** Date TBD around reading week. If you can't participate for medical or other reasons beyond your control, the weight of the mid-term oral is transferred to the final oral exam.
  - Individual oral exam on Zoom lasts 15 minutes with four questions. You are graded on the best 3 question, i.e. you can choose to skip one of the four questions.
- **Final oral: 10%** Last week of class.
  - Same format as the mid-term oral.
- **Project: 25%** You will work on a team project in which you solve a real problem involving real data and prepare a report including analyses, graphical displays and a careful interpretation of your analysis. The project has five components:
  1. A description of your plans including the data you plan to use and the general questions and methods of analysis you plan to use.
  2. An interim report on your progress submitted in early March, which your team will discuss with the instructor to get feedback.

3. A '.R' script using Markdown that produces a detailed analysis and presentation of your work, including diagnostics, etc. This output can be quite detailed.
4. A '.R' script using Markdown that produces an attractive and readable report with your main findings prepared in a way that would be suitable for a publication. You need to include all relevant references, data sources, etc. Aim for a maximum of 30 pages.
5. Slides for a **10-minute** presentation discussed below. The slides should be prepared with R-markdown using the ioslides format or other slide format. **You will collaborate using R, R Studio, R Markdown, git and github.**
6. You will prepare a brief summary of your project for a 10-minute presentation in late March. The 10-minute limit is strict. Be aware that it takes careful preparation and rehearsing to give a good presentation in such a short time. You must rehearse as a group ahead of time. The presentation will be followed by a 5-minute question and discussion period.

– The grade is based on the overall quality of the project (10%) and on your personal contribution to it (10%) and on your understanding of the issues and concepts in the project as shown in the final presentation and in project meetings with instructor. (5%).

- **Assignments: 20%**
  - Combination of individual and team assignments. Assigned approximately weekly. Most are done on Piazza. Some will involve contributions to Github R repositories.
  - Some assignments may have a higher weight than others.
  - All team members should feel responsible for helping each other to prepare and understand all solutions.
  - For team assignments, you split the questions evenly among team members and decide which team member will be responsible for taking the lead for which question(s). This is best done by agreeing to cycle through the questions in a systematic way. I will provide random numbers to help.
  - Team assignments are done in three steps. Usually, for an assignment given on Friday:
    * **Step 1:** to be completed by **deadline #1**, usually the following **Thursday at noon**:
      · The team member reponsible for a question posts a tentative solution on Piazza before **deadline #1**.
      · It must have a title of the form specified for the assignment.
      · The solution must start by **repeating the question** so someone looking at the solution can tell what question it solves.
      · For math, use the LateX editor in Piazza. You can also make sketches on paper, photograph them and upload the photograph to Piazza. Use markdown in R as much as possible.
      · When you first submit the post, make it **private to your team** and use the folder **assn X**, where **X** is the number of the assignment.
      · Each post remains **private** to your team until after **deadline #3**.
      · You get full marks for effort in making an honest attempt, it does not have to be completely correct.
    * **Step 2:** to be completed by all teammates by **deadline #2**, usually the following Saturday at noon:
      · Provide feedback on the solutions posted by your teammates: suggestions for improvements, improving coding in R, pointing out inconsistencies or errors, broadening the answer to cover a broader range of cases, etc.
    * **Step 3:** to be completed by **deadline #3**, usually the following Sunday at noon:
      · The team member responsible for a question reviews the suggestions made by teammates and incorporates them into the answer before **dealine #3**. Only **after deadline #3** and **before the next class**, make the solutions public to the class.
    * **Step 4:** Update your LOG file with links to the questions for which you took the lead and questions for which you provided help or comments. Add a line of the form `Assignment X: @123 @124 Comments: @111 @132` where `@123`, etc. are links to Piazza posts.
    * I will select some solutions as interesting sample solutions and add them to the **star** folder. Being added to the star folder does not necessarily imply that a solution is correct, nor does it mean that it's the best solution. It just means that I found some aspect of it interesting and

illustrative of the issues presented in the question. Conversely, not getting a star does not mean that you don't have an excellent solution. Sometimes you can learn as much or more from a solution with 'errors' than from a perfect solution.

- **Class and Piazza contributions: 10%** (possibility of bonus marks for outliers)
  - Contribute actively in class and post at least weekly on Piazza:
    * participate on Zoom responding and asking question orally or through the chat window.
    * post or edit **questions** and provide **answers** about course material
    * contribute to the course **wiki**: by posting your comments and/or links to something on the web that is interesting and relevant to statistics and add a link to it on the wiki page with a brief summary of the content and relevance
    * edit and improve existing posts
    * On or before February 15, add links in your LOG file to your best 5 contributions made before reading week and on or before April 4 add links to your best 5 contributions made after reading week. Add a line of the form: `Contribution: @41 @105 @ 107 @201 @261`
- **Optional weekly feedback every Friday evening and quiz questions: 5%** Every Friday starting with the third week) create a post that is private to the Instructor (it may be made public later during the weekend or you can make it public yourself) with information on each of the following:
  - What was the most interesting idea during the week?
  - What questions are you left with?
  - What was hardest to understand?
  - A quiz question based on the material of the week.
  - Add a link to your feedback in your **LOG** file.
- If you miss or are late for a component of the course for a medical, compassionate or technical reason beyond your control, the weight of that portion of the requirements for the course will be transferred to the final oral examination. Enter the fact that you missed and the reason in your LOG file so your grade can be computed correctly.

Prerequisites

The prerequisites for taking this course are MATH 3330, MATH 3131 and MATH 4330. Since MATH 3330 is a prerequisite for MATH 4330, we only check student records for MATH 4330 and MATH 3131.

References

- John Fox (2016) *Applied Regression Analysis and Generalized Linear Models, Third Edition*, Sage.
  - Web page
  - Online appendices
- Michael Evans and Jeffrey Rosenthal (2009) *Probability and Statistics – The Science of Uncertainty, 2nd ed.,* available online
- Hadley Wickham (2014) *Advanced R*
- Hadley Wickham (2015) *R Packages*

Getting Help

- Post questions and comments about the course material on Piazza. Post your questions to the entire class so everyone can benefit from the discussion and answer. I will monitor Piazza and participate if other students don't have an answer.
- If you have a personal question for the instructor, you can post it on Piazza as a private posting. This should only be used for personal questions that are of no interest to the rest of the class.
- If you happen to post a private question whose answer is of general interest to the class and that contains no personal information, I will assume that you consent to it being posted to the whole class unless you explicitly request otherwise.

- You can ask your teammates or other classmates directly.
- You can see the instructor during office hours or after class.

Some reflections on teams

The project and many activities are done in semi-randomly assigned teams that will be assigned during the first few weeks of class.

Working with a diverse team that you didn't select yourself gives you the opportunity to have experiences that will give you great anecdotes to use in your future job interviews.

When you land the job, you will be much more likely to show the kind of leadership and productivity in team work that is invaluable in the modern workplace.

Once teams are assigned, you will be able to communicate directly with your team by posting messages on Piazza and directing them to your team.

The more work you do on an assignment the better prepared you are to do well on the term test and on the final exam. But you shouldn't hog the work – let others do their part too. Everyone should make sure that they understand the whole assignment. Discuss the assignment with your team members to make sure everyone understands the key points and difficulties of each question.

Course policies

Missed deadlines

Late activities or projects will have their weight transferred to the final oral test.

Missed term test

If you miss the oral term test, the weight of the term test will be transferred to the final oral test.

Academic honesty

Familiarize yourself with the York University Senate Policy on Academic Honesty. Violations of academic honesty are treated very seriously in university. **Always cite your sources** for any information you use. This can as simple as providing links to websites you have visited to get information.

Bibliography

Evans, Michael J, and Jeffrey S Rosenthal. 2009. *Probability and Statistics: The Science of Uncertainty*. Second. Macmillan. http://www.utstat.toronto.edu/mikevans/jeffrosenthal/book.pdf.

Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. 3rd ed. Sage Publications.

Fox, John, and Jangman Hong. 2009. "Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package." *Journal of Statistical Software* 32 (1): 1–24. http://www.jstatsoft.org/v32/i01/.

Fox, John, and Sanford Weisberg. 2019. *An R and S-Plus Companion to Applied Regression*. 3rd ed. Sage Publications.

Monette, Georges, John Fox, Michael Friendly, and Heather Krause. 2018. *Spida2: Collection of Tools Developed for the Summer Programme in Data Analysis 2000-2012*.

Murnane, Richard J, and John B Willett. 2010. *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Snijders, Tom A. B., and Roel J. Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, Second Edition*. Sage.

Wickham, Hadley. 2014. *Advanced R*. CRC Press. http://adv-r.had.co.nz/.

———. 2015. *R Packages*. CRC Press. http://r-pkgs.had.co.nz/.