Course Description

MATH 4939: Statistical Data Analysis using SAS and R – Winter 2024

Georges Monette

January 2024

Contents

(Updated: January 07 2024 22:57)

This course description is tentative – final version available before January 19

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

Goals of MATH 4939

MATH 4939 is a fourth-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, imaginatively, creatively and correctly to the task of addressing real scientific or business questions with real data.

When you successfully 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 the design of a study, 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 is not apparent by merely inspecting 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: 10%**
  - Bi-weekly 10-minute quizzes every second Wednesday starting on January 17.
  - 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 exam.
- **Mid-term test: 20%**
  - In-class closed-book 50-minute midterm on Wednesday, February 14. There will be no quiz on this date.
  - If you can't write the midterm for medical or other reasons beyond your control, the weight of the midterm is transferred to the final exam.
- **Final exam: 25%** Written closed-book 2-hour exam during the regular exam period.
- **Project: 20%** 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 four 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 mid-March, which your team will discuss with the instructor to get feedback.
  3. A '.R' or '.Rmd' 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' or '.Rmd' 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 Rmarkdown 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 or early April. 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 (5%) 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: 15%**
  - Combination of individual and team assignments. Assigned approximately weekly or bi-weekly. Most are done on Piazza. Some may 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.
  - In most team assignments, different members of the team take the lead on preparing the answer to different questions.
  - These 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 Rmarkdown 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, your answer 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 **deadline #3**. Only **after deadline #3** and **before the next class**, **make the solutions public to the class**.
    * 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: 5%**
  - Contribute actively in class and make an original contribution at least **every two weeks** on Piazza (i.e. posting 7 items the last week doesn't count!):
    * 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. Include your **summary and critique** of the content and relevance. Use the **wiki** folder for these posts.
    * Edit and improve existing **wiki** posts.
    * At the end of the course, I may ask you to send me a post with links to your top 3 to 5 contributions.
- **Weekly feedback every Friday evening and quiz questions: 5%** Every Friday create a post that is private to the Instructors (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 or challenging idea during the week?
  - What questions are you left with?
  - A quiz question based on the material of the week.
  - Be sure to choose the 'Feedback' folder when you post your feedback.
- 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 examination.

Prerequisites

The prerequisites for taking this course are MATH 3330, MATH 3131 and MATH 4330. You must have passed **all** of these courses **before** taking MATH 4939. If you have deferred standing in one of these courses which you took in the fall term of 2023, you may enrol in this course but you must have passed all prerequisite courses to get credit for this course.

If you didn't take the prerequisite courses at York but are using courses taken elsewhere to fulfill prerequisites, you must contact me, preferably with a private message through Piazza, to schedule an appointment to test whether the material you have covered is an adequate equivalent. You may be required to pass a short test on the prerequisite material.

Textbook

This text and its online appendices are used mainly for reference. The textbook is the same as that used in the fall term for MATH 4330.

- John Fox (2016) *Applied Regression Analysis and Generalized Linear Models, Third Edition*, Sage.
  - Web page
  - Online appendices

References

- 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. The course's instructors will monitor Piazza and participate if other students don't have an answer.
- If you have a personal question for an 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 it contains no personal information, We will assume that you consent to it being posted to the whole class unless you explicitly request otherwise. When such posts are posted to the whole class, the original author remains anonymous unless you identify yourself in the post.
- You can ask your teammates or other classmates directly.
- You can ask instructors during tutorials, office hours, or after class.

Some reflections on teams

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

One of the major focuses of most job interviews is your skill at working in teams. 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.

Many employers prefer applicants who will not only be productive themselves but also contribute to the creativity, imagination and productivity of their colleagues and teams.

Common interview questions stress responsibility and cooperation. They might go along the lines of:

- Tell us about a team project you worked on. How did you help other team members achieve their potential?
  What was your role and what did you contribute?
- Tell us about a time you had to work with a colleague who was difficult to get along with.
- Tell us about a time your team didn't want to use your idea, what did you do and what was the outcome?

Interviewers are not looking for a 'right answer'. They are giving you a chance to show whether naturally think of helping others achieve their potential, in addition to being highly competent yourself.

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 workplace.

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

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

In general, late activities or projects will have their weight transferred to the final exam.

Missed term test

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

Academic honesty

In **this course**, you are allowed to use aids like ChatGPT or other sources that **may be forbidden in other courses.** When you have used such sources, cite and **credit** their answers in your answer, and include an analysis and critique of their answers. Much of the material of this course involves recognizing fallacies and paradoxes in statistics and probability. Fallacies are commonly-held but incorrect beliefs that ChatGPT probably shares. In my experience, ChatGPT provides lengthy rambling answers that include excellent insights along with serious errors and superficial irrelevant commentary. Since you will be using these tools in your future work, developing critical experience with them is valuable. For that reason, you may use any source you wish in your work, **provided the material and its source is cited and that you include your own critical analysis of the material.**

It is vital that you familiarize yourself with the policies of each course. If you are in doubt, ask the instructor. In most courses, any use of ChatGPT or other similar sources is considered a breach of academic honesty and may result in serious disciplinary consequences.

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. Doing so enhances the value of your work.

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." https://github.com/gmonette/spida2.

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/.