Methods of Data Analysis II

University of Toronto
Department of Statistical Sciences
STA303H1S/1002HS Summer 2021

LEC 0101

Instructor: George Stefan
Email: george.stefan@mail.utoronto.ca

Class Day/Time: MW 9AM–12PM EDT
Office hours: TR 4–5PM EDT on Bb Collaborate

---

* ***This is an online course. Please note that since lectures and/or evaluations will be taking place during the above lecture times, you must be available during those times. No accommodations will be made for assessments missed during these times.***

** ***As this is an online course and all assessments must be submitted through Quercus, it is the STUDENT'S responsibility to ensure they have a reliable internet connection.***

COURSE OVERVIEW

Course Description: Briefly, the contents will focus on the exponential family, generalized linear models (GLMs), quasi-likelihood, generalized estimating equations (GEEs), linear mixed models, generalized linear mixed models (GLMMs) and generalized additive models (GAMs). Statistical analysis will be conducted using `R`. The students will be expected to simulate datasets using `R`, as well as interpret `R` code and output on tests and projects.

Content, emphasis, etc. of the course is defined by means of the lecture material — not only the posted lecture notes. It is important to attend all lectures, as there is normally no simple way to make up for missed lectures (perhaps obtain another student's notes). There will also be many lecture examples using statistical software `R`.

Learning Outcomes: By the end of this course, all students should have a solid understanding of both the mathematical theory of GLMs, LMMs, GLMMs and GAMs, as well as their application in the form of 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 GLM concepts and theory;
- practical application of methods on real data using statistical software `R`, with appropriate justification of use of these methods;
- interpretation of data analysis results in clear and non-technical language

Pre-requisites: The prerequisites are:
Prerequisite: STA302H1 or STAC67H3 or STA302H5 or STA1001H
Exclusion: STAC51H3
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.

COURSE MATERIALS

Course Content: All lecture slides, recordings and materials will be posted on the Quercus course page for each lecture section. Furthermore, any important announcements will also be posted in Quercus. Please make sure to check it regularly so you don't miss anything.

Textbook: There are no required textbooks. All assessments will be conducted based on lecture materials. However, there are some useful books which can be used for references.

1. Alan Agresti. "*Categorical Data Analysis (3rd edition)*". Wiley, 2011.
2. David Clayton and Michael Hills. "*Statistical Models in Epidemiology*". OUP Oxford, 2013.
3. Peter McCullagh and John Ashworth Nelder. "*Generalized Linear Models (2nd Edition)*", Chapman & Hall/CRC, 1989.
4. Peter Diggle, Patrick Heagerty, Kung-Yee Liang, Scott Zeger. "*Analysis of Longitudinal Data*". Oxford University Press; 2002
5. Julian Faraway. "*Extending the Linear Model with R: Generalized Linear, Mixed Effects and Non-parametric Regression Models*". CRC press, 2016.
6. Simon N. Wood. "*Generalized Additive Models: An Introduction with R*". CRC press, 2017.

Statistical Software: We will be using RStudio to perform 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 rstudio.cloud (**link**) which only requires you to login with your UToronto email and connect to our course project via the link provided. Support for downloading and learning R (and RStudio/RCloud) will be provided during lectures or through documents on Quercus. In lectures, examples with R syntax will be provided, which should be sufficient for you to learn how to apply the statistical methods.

COURSE COMPONENTS

Lectures: Lectures will take place live on Bb Collaborate through Quercus with recordings posted afterwards. During lectures and videos, we will cover important course materials, as well as cover a number of examples illustrating the uses of these methods. Lecture slides/videos will contain some `R` code and output to show how to perform these methods in practice. Each lecture builds on the material from previous weeks, so it is recommended that you attend lectures regularly/stay on top of the material.

Office Hours: Instructors and TAs will hold office hours through Bb Collaborate in the Quercus course page. The office hour schedule will be posted on Quercus. 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 accelerated class to have material clarified as quickly as possible. Don't wait until the last minute to ask your questions!

Quercus Discussion Board: We will be using the Quercus Discussion Board as an online discussion forum. **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.

GRADING SCHEME

All students will be evaluated in the following way:

Assessment	**Date Due/Occurring**	**Marks (%)**
"Weekly" Online Quizzes (×5)	At the end of each week period (from week 2)	25%
Discussion Board Participation	Ongoing	5%
Term Test	July 26 from 9AM–12PM EDT	20%
Assignment 1	July 23 at 11:59PM EDT	10%
Assignment 2	August 13 at 11:59PM EDT	10%
Final Project	August 22 at 11:59PM EDT	30%

Please note that the last day to drop the course without penalty is August 2, 2021.

EVALUATION BREAKDOWN

"Weekly" Online Quizzes: There will be 5 "weekly" online quizzes, each worth 5% of the overall grade; these will be occur during the last 30 minutes of the lecture time of each Wednesday. Quizzes will begin on **Wednesday, July 14** and continue until the last lecture period. The quizzes will be uploaded on Quercus, so **please make sure you have a decent internet connection**.

- The quizzes will be multiple choice and cover material from the previous set of lectures. You may wish to have a calculator/computer available at this time to aid in any calculations/computations.
- Quizzes can be found on Quercus under the "Quizzes" tab in the navigation bar, or through the link provided in that week's module, and will only be available during the designated quiz time. Quizzes must be done individually.
- **Missed quiz policy:** Students can miss up to **two (2)** quizzes without academic penalty. The student will need to provide **24-hour notice** to the instructor; i.e. if a quiz is scheduled for Wednesday at 8:30 pm EDT, the student must send an email before 8:30pm on Tuesday. Otherwise, the quiz will count and a zero will be assigned. If a quiz is missed and advance notice is given, the remaining quizzes will be evenly reweighted and will still be worth 25% of the overall grade. For example, if one quiz is missed, the remaining four quizzes will each be worth 25/4=6.25%; if two quizzes are missed, the remaining three quizzes will each be worth 25/3=8.33%. Once a student decides to take a quiz, the grade will be counted.
- **All students must attempt at least three (3) quizzes to pass the course.** No further accommodations will be made for missed quizzes.

Term Test: The term test on July 26 has to be submitted online via Quercus. The term test will be based on the contents of the first six lectures. The focus of the term test will focus on the mathematical aspects of GLMs. More information on this test will be provided later.

Assignments: You will be assigned two assignments in the term. The purpose of these assignments is to develop your theoretical and data analysis skills which will be useful for the final project and future

courses. The assignments will have a strong focus on the use of statistical software (`R` specifically), and will involve applying the methods learned during lecture to a dataset. **Late assignments will be penalized**. Late submissions will receive a 20% penalty for each day that the project is late. In general, extensions will not be given unless a valid reason is provided. In such cases, the instructor may decide to grant an extension of up to 5 days.

Final Project: The final project will be due on **August 22, 2021 by 11:59PM EDT** and will consist of data analysis on a novel dataset. Students will be required to demonstrate their understanding of the methods taught in lecture by developing a reasonable regression model 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 will be submitted as a project report, which comprises:

- Introduction section: provides details regarding why the model is being developed, general information regarding how the model is developed and finally how the model meets the purpose mentioned earlier.
- Exploratory data analysis section: a detailed description of the variables in the data with appropriate tables or figures that highlight certain characteristics deemed relevant or important.
- Model development section: a detailed discussion of the process used to come to the final model, as well as in-depth diagnostics to illustrate the 'goodness' of the model.
- Conclusion section: restate why the model is useful in the context of the data, provide an interpretation of the final model in non-technical language, and discuss any limitations/problems remaining with the model and how they might impact its use in the real world.

The final project will be an individual project, and must be typed and submitted by the stated deadline. A word count limit will be given, as well as other more detailed instructions at a later date. **In order to pass the course, you must submit the final project.**

MISSED ASSESSMENT POLICY

Students are responsible for completing all of the assessments detailed in the previous section. If a student is sick and needs to request an extension or accommodation on a mini project, they must send an email to their instructor. In order for the request to be considered, the email:

- must be received at least one day before the mini project is due;
- must include the course code in the subject line;
- must include your full name and student number;
- must specify for which project the extension/accommodation is being requested;
- must include the following sentences:
  - "I affirm that I am experiencing an illness or personal emergency and I understand that to falsely claim so is an offence under the Code of Behaviour on Academic Matters."
  - "I understand that the weight of this assessment will be moved to the weekly quizzes (10%) and to the final project (5%)"

In order to pass this course, students must submit the final project, at least one of the assignments or the term test and have attempted three (3) of the quizzes.

COMMUNICATION

Please do not email the instructor with questions related to the content of the course. These types of questions are much easier to answer through the discussion board or during office hours. Emails that do not contain sensitive or personal information will be directed to post the questions on the discussion board. If you need to email the instructor for personal reasons, please use your official University of Toronto email address, include STA302 in the subject and also include your full name and UTORid in the body of the email (in case we need to look anything up).

INTELLECTUAL PROPERTY

Course materials provided on Quercus, such as lecture slides, assignments, tests and solutions are the intellectual property of your instructor and are for the use of students currently enrolled in this course only. **Providing course materials to any person or company outside of the course is unauthorized use**. This includes providing materials to predatory tutoring companies.

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:

- You may consult class notes/lecture slides during quizzes and tests, 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.
- Paying anyone else to complete your assessments for you is academic misconduct.
- Sharing your answers/work/code with others is academic misconduct.
- Looking up solutions to test/quiz problems online or in textbooks and copying what you find 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 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.

CLASS SCHEDULE – TENTATIVE

This is the tentative outline for Summer 2021. Topics may be reduced or additional topics may be added by course instructor's discretion.

Lecture	Content
1 (July 5)	**Introduction:** Review of categorical data analysis for univariate and bivariate models.
2 (July 7)	Study designs, contingency tables, risk difference and risk ratio. Odds ratio, rate ratio, delta method, confounding and interaction. Test of independence.
3 (July 12)	Small sample inference. Fisher's exact test. Review of linear regression. Introduction to GLMs. Logistic regression. Exponential family and iteratively reweighted least squares (IRLS)
4 (July 14) Quiz 1	Poisson regression. Overdispersion and negative binomial regression. Offset terms in GLM. Multinomial GLM, proportional odds GLMs and probit GLM.
5 (July 19)	Model diagnostics. Classification and Discrimination using Logistic Regression. Cross validation, ROC curve and Bootstrap method.
6 (July 21) Quiz 2	Variable selection: Stepwise methods, LASSO. Matched case control and conditional logistic regression.
7 (July 26)	**Term Test**
8 (July 28) Quiz 3	Linear mixed effects models (LMM)
August 2	Deadline to drop course without penalty
9 (August 4) Quiz 4	Linear mixed effects models (LMM)
10 (August 9)	Generalized linear mixed effects model (GLMM)
11 (August 11) Quiz 5	Quasi Likelihood and generalized estimating equation (GEE)
12 (August 16)	Generalized Additive Models (GAM) and Generalized additive mixed models (GAMM)
August 18–30	Final assessment period