STA303 / 1002 HF - Methods of Data Analysis II

Jeffrey Negrea

Summer 2017

Course description

This course extends the linear model from STA302 (methods of data analysis I) to include indicator variables, correlated errors and link functions. Topics to be covered include: Analysis of Variance for one- and two-way layouts, logistic regression, loglinear/Poisson regression, longitudinal, repeated measures and mixed models, and non-linear regression. This course will also be an opportunity to continue to develop skills in data analysis for which the **R** software and **Rmarkdown** will be used.

Pre-requisite

Students should have STA302 or equivalent preparation. Students are also expected to have the mathematics pre- and co-requisites required by students in all courses leading up to STA302. This course is slightly less theoretical than STA302, but please do not attempt the course without the required mathematical background.

Lectures

- Monday 9:10-12:00 in **MP102**
- Wednesday 9:10-12:00 in **MP102**

Important Dates

First Day of Classes	Jul. 5
Last Day to Add	Jul. 10
Last Day to Drop	Jul. 31
Civic Holiday (University Closed)	Aug. 7
Last Day of Classes	Aug. 14
Exam Period	Aug. 15 - 18

Instructor & TA Office hours

- Instructor: **Jeffrey Negrea** (jeffrey.negrea@mail.utoronto.ca). Office: SS 6025
- Office hours (an ideal time to discuss questions that you have)
  - Instructor: Mondays and Wednesdays 2:00-3:00 in SS 6025
  - TA office hours will be scheduled before midterm and assignments due dates. Check out announcement on portal.

In general, I prefer not to answer questions about the course content by e-mail. Questions about the course content should be addressed on the Piazza discussion board or during office hours. Before you send an e-mail, make sure that you are not asking for information that is already on the course web site or the Piazza discussion board. Issues not relating to course content can be resolved by email; please put **STA303:** at the start of the subject line.

Announcements will be posted on Blackboard. Please check portal regularly. If an urgent matter arises, I may contact the entire class by e-mail. In order to receive these message, please make sure you that you use your mail.utoronto.ca account so that the message won't automatically go to the Junk folder.

Course website

Weekly lecture notes, assignments, practise problems, and announcements are available on

https://portal.utoronto.ca

Please note that we have a **Discussion board** on Piazza, which I will be monitoring. If you post your questions there and don't get response in 4 days, please send me an email. Here is the signup link:

https://piazza.com/utoronto.ca/summer2017/sta303sta1002/home.

Textbooks

We don't have a specific textbook for this course. Here is a list of recommended references.

- *KNN: Applied Linear Regression Models*, 4th edition by Kutner, Nachtsheim, and Neter. (We will be covering most of Chapters 8, 11, 13 and 14. This is a good textbook and worth the read, although it is not required for the course.)
- *SJS: A Modern Approach to Regression with R* by Simon J. Sheather. (It is currently available online (as an e-Book) through the library website. We will be covering material from Chapters 4, 8 and 10.)
- *SW: Applied linear regression* 4th edition by Sanford Weisberg.

Evaluation

Item	**Weight**	**Date**	**Time**	**location**
Assignment 0	2%	Monday, July 10th	Due: 10pm	**TBD**
Assignment 1	8%	Monday, July 17th	Due: 10pm	**TBD**
Assignment 2	10%	Wednesday, July 26 th	Due: 10pm	**TBD**
Midterm	25%	Monday, July 31st	In Class	In Class
Assignment 3	10%	Wednesday, August 9th	Due: 10pm	**TBD**
Final Exam	45%	TBA	3-hour exam	TBA
Contribution on Piazza	1 or 2 point(s)	add to Final Exam	April 10th	

The midterm and exam are both **closed book and closed notes**, a non-programming calculator is allowed. The midterm will be written in the lecture room and another booked room (location will be announced later). Practice problems will be posted on portal to help you prepare for the midterm and exam and are not to be handed in. Each assignment will mainly be a data analysis project for which you will use R.

If the midterm is missed for a valid reason, you must provide appropriate documentation, such as the University of Toronto Medical Certificate, University of Toronto Health Services Form, or College Registrar's Letter. You must submit this documentation within one week of the test. If documentation is not received in time, your

test mark will be zero. If midterm is missed for a valid reason, the weight of the midterm will be shifted to the Final Exam. Your midterm mark will be zero if you miss it for an invalid reason. **The midterm is a 2-hour test and scheduled during class time, please enrol the course only if you are able to write the midterm.**

Assignments submitted late without a valid reason will get penalty of **10%** per day late.

Any requests to have marked work re-evaluated must be made within 7 days (one week). The request must be accompanied by an email which explains what was marked incorrectly. **All remark requests, the whole assignment/ test will be remarked.** There is chance that you might get lower mark points after remarking.

Computing

We will be using R and R-Studio. The main advantages of R are the fact that R is freeware and that there is a lot of help available online. For instructions on how to install R and RStudio, and to learn the basic syntax of R, refer the documents created by Paul Torfs & Claudia Brauer

https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

I am assuming that students have used R before. Note that there are many graphics options available to produce the sophisticated plots that are in the book or online, but we will focus on the basics. There are many good reference online, if **R** is new to you, here is another 100 page document which may be helpful:

http://www.utstat.toronto.edu/~brunner/help/R-intro.pdf

For assignments, you will use Rmarkdown or LaTeX to write your solution (PDF submission is preferred). For the first assignment, I will provide you an Rmarkdown template and a LaTeX template. To learn more about Rmarkdown, refer to

http://rmarkdown.rstudio.com

First time user of R markdown:

https://www.youtube.com/watch?v=QaKCirYknS8

RMarkdown short tutorial with RStudio:

https://www.youtube.com/watch?v=DNS7i2m4sB0

Academic Integrity

It is an academic offence to present someone else's work as your own, or to allow your work to be copied for this purpose.

Here are some guidelines that apply to the computational assignments.

- In this course, it is always okay to use computer code that is presented in lecture or the textbook. Use it any way you like; you are responsible for the results.
- *The biggest danger is copying from other students in the class.* It is fine to discuss the assignments and to learn from each other, but don't copy.
- Above all, do not allow anyone to see your program file before an assignment is due, and do not look at anyone else's. Never photograph someone's solution or allow yours to be photographed. To repeat: **the person who allows her/his work to be copied is equally guilty, and subject to disciplinary action by the university.**
- It is acceptable to get help with your assignments from someone outside the class, but the help must be limited to general discussion and examples that are not the same as the assignment. As soon as you get an outside person to actually start working on one of your assignments, you have committed an academic offence.
- *Don't copy, and don't let anyone copy from you.*

- If this is not clear enough, the latest version of the student handout "How not to Plagiarize" is available at http://www.writing.utoronto.ca/advice/using-sources/how-not-to-plagiarize

  You are responsible for knowing the content of the University of Toronto's Code of Behaviour on Academic Matters at http://www.governingcouncil.utoronto.ca/policies/behaveac.htm

If you have any questions about what is or is not permitted in this course, please do not hesitate to contact me. It is OK to discuss assignment problems with other students in the class or discussion board on Piazza. However, assignments must be written up completely by yourself. Do not let other students read your completed assignment solutions as this can lead to copying. Failure to comply with this is a serious academic offence.

Course Schedule

The tentative schedule, as of June 29, 2017, of STA303/1002 course follows. The schedule may change as circumstances necessitate.

Lecture	Reference	Topics	Notes
1	KNN: Ch8 SJS: CH1	Introduction to course t-tests 1-way ANOVA	Try first R markdown lab: test.pdf -> 3 pdf files
2	SJS: Ch 2	2-way ANOVA	A0 due
3	KNN: Ch 2 SJS: Ch 2	ANCOVA	
4	KNN: Ch 11.1 SJS: Ch 4	Weighted Least Squares (WLS) Regression	A1 due
5	KNN: Ch 11.2	Ridge regression	
6	KNN: Ch 14.1-14.4 SJS: Ch 8	Logistic regresssion	
7	KNN: Ch 14.1-14.4 SJS: Ch 8	Logistic regression	A2 due
8		Logistic regression with replicates	Midterm
9	KNN: Ch 14.13 SJS: Ch 4	Poisson regression	
10	SJS: Ch 5	Log-linear model for count data	
11	SJS: Ch 10	Repeated Measures ANOVA Linear mixed effect model	A3 due
12		Linear mixed effect model	