STA302/1001, Autumn 2017
Methods of Data Analysis I

Instructor: Professor Mark Ebden
Email: mark (dot) ebden (at) utoronto (dot) ca
Office: SS 6027 CLTA
Office hours:
Section 1: From 1:10 to 2:50 pm on Wednesdays, by appointment (click here)
Section 2: From 3 to 4 pm on Wednesdays, by appointment (click here)

Teaching Assistants:

Section 1:	Xing Shuo Zhai	Section 2:	Lu Yu
	Lei Sun		Debby (Shih Yung) Hwang
	Boris Garbuzov		Alex (Yuxiang) Gao
	Brian (Xin) Ning		Pip (Phillipa) Swartz

TA Office Hours:

- Mondays 4:30-6 pm with Brian, starting 18 September
- Tuesdays 5-7 pm with Pip or Alex (a time-share setup), starting 19 September
- Fridays 10-11:30 am with Lei, starting 15 September

Course webpage: Accessible through the learning portal: portal.utoronto.ca

Classroom sessions:

- **Section 1:** Twenty-three sessions (excluding midterm) from Thursday 7 Sept to Tuesday 5 Dec; in room OI G162, 10-11 am (Thurs) & 10 am – 12 pm (Tuesdays)
- **Section 2:** Eleven sessions (excluding midterm) from Thursday 7 September to Thursday 30 November; in room ES 1050, 5-8 pm

In each section, the aim is to have one break per week of at least 10 minutes during.
A link to our campus map is here.

Prerequisites

- Required: STA248H1/STA255H1/STA261H1/ECO227Y1
- Suggested: CSC108H1/CSC120H1/CSC121H1/CSC148H1
- Suggested: MAT223H1/MAT240H1

You will need to recall basic matrix operations which you learned in the linear algebra course. An example of a review of matrix algebra is available in the first ten pages here.

Course Content

This course covers the theory and application of linear regression analysis. In particular this course will cover: initial examination of data, correlation, simple- and multiple regression models using least squares, geometry of least squares, inference for regression parameters for normally distributed errors, confidence and prediction intervals, model diagnostics and remedial measures when the model assumptions are violated, interactions & dummy variables, ANOVA, model selection, and ridge regression.

The learning objectives of this course are:

- To gain a solid understanding of linear regression analysis
- To learn practical skills for developing linear regression models for inference and prediction, and to interpret the results

Textbook

A Modern Approach to Regression with R by Simon Sheather (2009)
- ISBN: 978-0387096070
- Most of chapters 2-7
- The textbook website has datasets and other resources

Additional reference material (optional)

Applied Linear Regression Models, 4th edition, by Michael Kutner, Christopher Nachtsheim, and John Neter (2004)
- ISBN: 978-0073014661
- Much of Chapters 1-8, and some content from chapters 9-11, overlaps with our course

Applied linear regression, 4th edition, by Sanford Weisberg (2014)
- ISBN: 978-1118386088

Evaluation

Undergraduate and graduate students will be evaluated according to the following marking scheme.

	Weight	Dates	Time
Term test	40%	Section 1: Tue 24 Oct Section 2: Thu 26 Oct	During class
Final exam	60%	Expected 9-20 Dec, same time for 2 sections	Scheduled by Faculty

Graduate students will be evaluated at the graduate level according to the University Assessment and Grading Practices Policy.

Term test and exam

You must bring your student identification to the term test and the final exam.

The midterm and exam are both closed book and closed notes. A non-programmable calculator is allowed. The location of the midterm will be announced later. The midterm papers may be different for sections L0101/L5101 but the final exams are identical. Practice problems will be posted on Portal to help you prepare for the midterm and exam and are not to be handed in. You need to know basic R syntax for the midterm and exam, and you must also know how to interpret output from R.

Marking concerns

Any requests to have your test mark re-evaluated must be made in writing within *one week* of the date the work was returned. The request must contain a justification for consideration.

Missed Tests

- If a test is missed for a valid reason, you must submit documentation to the course instructor.
- If a test is missed for a valid medical reason, you must submit the University of Toronto Verification of Student Illness or Injury form to your instructor within one week of the test.
- The form will only be accepted as valid if the form is filled out according to the instructions on the form.
- Important: The form must indicate that the degree of incapacitation on academic functioning is moderate, serious, or severe in order to be considered a valid medical reason for missing the term test. If the form indicates that the degree of incapacitation on academic functioning is negligible or mild then this will *not* be considered a valid medical reason.
- If a test is missed for a valid reason then the final exam will be worth that much more of your final grade.
- Other reasons for missing a test will require prior approval by your instructor. If prior approval is not received for non-medical reasons then you will receive a term test grade of zero.

Computing and Calculators

For the test and exam, you will need a basic scientific hand-calculator, with statistical functions, logarithmic functions etc. Calculators on phones or other devices equipped to

communicate with the outside world (for example, through the internet or cellular or satellite phone networks) will not be permitted during the term test and the final exam. Programmable calculators are not allowed on the midterm or exam.

We will make extensive use of the R language (and RStudio) to analyse data. The main advantages of R are that it is freeware and that there is a lot of help available online. R is freely available for download at cran.r-project.org for Windows, Mac, and Linux operating systems. *RStudio* is a good integrated development environment to R. It is freely available at www.rstudio.com/products/rstudio You may also like to sign up for a CQUEST account. To get an account and find out more information about using CQUEST go to www.cquest.utoronto.ca

If you need help installing R and RStudio, and learning the basic syntax of R, a helpful document is here. Another good online R reference is here, and there is a downloadable book called *Introduction to R* by William Venables and David Smith. The alternative reference textbooks above, and websites such as datacamp.com, are also good to help you pick up R.

I'll give you a quick introduction to R and provide you with R source code for the examples in lecture. Note that there are many graphics options available to produce plots, but we'll focus on the basics, certainly sufficient for the (non-credit) assignments. You may like to use *Rmarkdown* to type up your work.

Online Discussion Board

You will have the option to use Piazza for class discussion. If you decide not to use Piazza it will not disadvantage you in any way, and will not affect official University outcomes (e.g., grades and learning opportunities). If you choose not to opt-into Piazza then you can ask questions or discuss course material with the instructor or TAs during office hours.

Please read Piazza's Privacy Policy and Terms of Use, taking time to understand and be comfortable with them. They provide for substantial sharing and disclosure of your personal information held by Piazza, which affects your privacy. If you decide to participate in Piazza, only provide content that you are comfortable sharing under the terms of the Privacy Policy and Terms of Use.

The Piazza system is highly catered to getting you help quickly and efficiently from classmates, the TA, and the lecturers. Rather than emailing questions to the teaching staff, we encourage you to post your questions on Piazza. To sign up for the discussion forum, click on the link: piazza.com/utoronto.ca/fall2017/sta3021001

TAs are assigned to answer questions you have on Piazza. If you post your questions there and don't get a response in three days, please inform me as soon as possible.

Additional help

Practice problems list from the textbook for you home preparation will be posted on the web-site. They are not to be handed in. The solutions will not be posted. You may discuss them in tutorials etc.

Need extra help with the coursework? Here are some options:

- For continued class discussion and questions outside of class, try posting on the discussion forums. The instructor and TAs will be monitoring the
- You may choose to join (or create) an STA302 Recognized Study Group: www.studygroups.artsci.utoronto.ca
- You can visit the instructor or teaching assistants during their office hours
- E-mail should only be used for emergencies or personal matters

How to communicate with your instructor

Questions about course material such as:

- How do I do question 3.7 in this textbook?
- What is standard deviation?
- When is the midterm?

can be posted on the Piazza discussion forum. If you are shy, questions can be posted anonymously (so that the author is anonymous to other students but not to the instructors).

For private communication, such as "I missed the test because I was ill," e-mail me. Use your utoronto.ca e-mail account and include your full name and student number.

You may post entirely anonymous feedback here. Nobody will receive this except me, and nobody will know who you are.

Academic integrity

You are responsible for knowing the content of the University of Toronto's Code of Behaviour on Academic Matters at 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 your instructor.

Accessibility needs

The University of Toronto is committed to accessibility. If you require accommodations for a disability, 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.

Your responsibilities

The classroom sessions for this class are designed to actively engage you in the course material. We hope you'll find them interesting, challenging, fun, and an excellent opportunity to truly learn the material.

Tentative Schedule

This is subject to change, with adequate notice.

- Week 1: Introduction
- Week 2: Chapter 2
- Week 3: Chapter 2
- Week 4: Chapters 2 and 3
- Week 5: Chapter 3 (end of midterm coverage announced)
- Week 6: Chapter 3
- Weeks 7-12: From Chapters 3 to 7, and ridge regression
    - In Chapter 3, we'll skip inverse response plots and the Box-Cox procedure in Section 3.3.3
    - In Chapter 4, we'll skip Sections 4.1.1 (PIs for weighted least squares), 4.1.2 (leverage for WLS), and 4.1.3 (using LS to calculate WLS)
    - For Chapters 5 to 7, details of skipped material will be announced