STA302H1F / 1001HF – Methods of Data Analysis I

Fall 2015

Lectures: L0101: T10-12, R10-11 in ES 1050
L0201: R5-8 in SS 2117

Instructor: Craig Burkett (burkett@utstat.utoronto.ca)
Office: SS 6015

Office Hour: Thursdays 3:00 – 5:00pm + additional hours around the tests.

Web-page: http://portal.utoronto.ca (U of T Blackboard)

TA office: SS 1091 M2-4, T1:30-3:30, W11-12, R11-12
During the office hours, a teaching assistant will be available to answer questions. Tests and assignments will be returned at these times.

Overview: This course covers some of the theory and methodology of data analysis when linear regression models are appropriate. Topics to be covered include: initial examination of data, correlation, simple and multiple regression models using least squares estimation, inference for regression parameters under assumptions of Normally distributed errors, confidence and prediction intervals, diagnostics and remedial procedures when model assumptions are violated, interaction and dummy variables, measurement error and model selection. This course will also be an opportunity to begin to develop skills in data analysis for which the R software program will be taught.

Textbook: The recommended textbook is *Applied Linear Regression Models 4e* by Kutner, Nachtsheim & Neter (ISBN: 9780073013442). We will be covering most of Chapters 1 through 8 and selected material from chapters 9 and 10, as time permits. This is a good textbook and worth the read, although it is not required for the course. I still have a copy of it on my bookshelf.

Another good textbook is *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 most of Chapters 1 through 7, excluding 4. Topics in later chapters will be covered in STA 303H1. This book was used previously in the course, and the notation is noticeably different so be careful if you read it. Datasets and other resources are available at the textbook website http://www.stat.tamu.edu/~sheather/book/.

Another text I considered is *Regression Analysis 7e* by Mendenhall & Sincich (ISBN: 9780321691699). It is noticeably simpler than the other two texts, but a good choice if you really don't know what's going on, or if you're coming into the course with very little mathematical background.

Prerequisites

Students should have a second year math-based statistics course such as STA 248H1 / STA 261H1 / STA 255H1 or ECO 227Y. Regardless of how you arrived here, I expect that you have knowledge of Appendix A (up to the end of A.7), for example. There is also a document posted on Portal (courtesy of A. Gibbs) for those who cannot see the Appendix.

Students are also expected to have the mathematics pre- and co-requisites required by students in these second-year statistics courses. You will need to know basic matrix operations. A good review of the matrix algebra that we will need can be found in the first 10 pages of this document.

http://www.stat.ncsu.edu/people/davidian/courses/st732/notes/chap2.pdf

Most applied courses in the Statistics Department require STA 302H1 as a pre-requisite. As a consequence, this course has a theoretical component to prepare students for more advanced work. Please do not attempt the course without the required mathematical background.

Follow-up courses

STA 303H1 (Methods of Data Analysis II) focuses on aspects of linear models that are not covered in STA 302H1 such as non-Normal and correlated response variables.

Evaluation

The grading scheme is as follows:

Assignment 1	5%	Due: October 15 (at start of lecture)
Mid-term Test	25%	L0101: Tues. Oct 20 @ 10am
		L0201: Thurs. Oct 22 @ 5pm
Assignment 2	10%	Due: November 5 (at start of lecture)
Assignment 3	10%	Due: December 3 (at start of lecture)
Final exam	50%	During exam period

If your exam mark is better than your term mark (including the exam), the exam weight will be 60% and the test weight will be 15%. The test room will be posted on the course website prior to the test. The assignments will involve a data analysis project for which you will use R. You will not need to know R syntax on the test and exam, but you will need to interpret output from R.

No late assignments will be accepted without documentation of a valid reason. Lateness penalties are at the discretion of the instructor.

STA 1001 students should speak to me regarding an optional adjustment to the marking scheme.

Practice Problems

Assigned practice problems are **not** to be handed in. They are simply for your own practice, for the tests and assignments.

Professor Contact

There are various ways in which the TAs and I would be happy to serve you. Here are some rough guidelines:

- If you have a personal issue that you believe I can resolve in a few minutes, please speak to me before or after lecture, or during a break. You can also come to office hours if you require more time or privacy.
- If you'd like to discuss the class material in more depth, please come to office hours. You can also try me after class or at a break, but priority will be given to above.
- If you'd like to discuss the solution to homework questions, please post on the discussion forum. If you don't get a satisfactory answer, please see the TA. They will probably be more familiar than me with specific questions. If you're not satisfied with their answers, please come to office hours.
- If you want to ask a question about the course content, a practice problem, an announcement that was made in class but you missed it because you were not present or not listening, please use the discussion forum on Portal.
- If you have an issue that must be dealt with by me, and can be handled in three sentences of text or less, or to report a problem with Portal or the assigned homework, or to inform me of something relevant to the course (such as a missed test), please send me an email.
    - If your email can be answered by reading this syllabus or the Portal discussion forum, I will not answer it. Please don't be offended.

NB: I don't check email constantly as, believe it or not, I don't have a mobile phone. I also teach several hundred students (~ 650 this term), and cannot handle the volume of emails that come through with that number. Further, I don't really like sending/receiving emails, and would much prefer that you speak to me in person. That said, if you believe an email is appropriate, please email me using your *@utoronto.ca or *@mail.utoronto.ca address. You won't get a response if you email from other email addresses, and it probably won't even be read since my spam filter may block it. The reason for this is so that I know whom I am writing to, and so that I don't provide any personal information to someone who shouldn't be receiving it. Also, please put "STA302: " at the start of your subject, as I teach multiple courses most terms.

Important Notes

- If a test 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 to the course Instructor within one week of the test. Print on it your name, student number, course number and date, and have the doctor record the reason for the visit. No notices will be accepted without a CPSO number stamped on the form (ie. they must be a doctor in the *western* sense of the word, not somebody who reads tea leaves).
- If documentation is not received in time, your test mark will be zero. If the test is missed for a valid reason, I reserve the right to force you to make up the test OR to shift the weight to the final exam, at my discretion. Most likely you will be writing a makeup test.
- Any requests to have marked work re-evaluated must be made **in writing** within one week of the date the work was returned to the class. The request must contain a justification for consideration; do not simply write "please see #3".
- The course website will be used to post lecture notes, R examples used in lectures, practice problems, assignments and solutions, past tests/exams, other course info and important announcements. **Check it regularly**. The website also has an electronic discussion forum that you can use to communicate with other students in the course.
- If an urgent matter arises, I may contact the entire class by email. In order to receive these messages, please make sure that your ROSI account has your **utoronto.ca** email.
- In general, I am not able to answer questions about the course material by email. Before sending an email, make sure that you are not asking information that is already on the course website, or questions about the course material or assignment that are more appropriate to discuss through the forum.
- Questions about the course material can be posted on the discussion board on Blackboard. Other students may be able to answer your questions very quickly, and the TAs will check the board regularly as well.

Computing

Historically, this course has been taught using SAS. This term marks the first time that we will be delivering the course using R Statistical Software, for various reasons. While many of you will be happy with this change, please note that all past tests and exams will have SAS output and not R, so you'll probably need to learn how to read both. At least you don't need to learn SAS coding!

You can download R for free at:
http://cran.r-project.org/
Don't forget to select the correct operating system! This site will give you a file to install base R on your system. Other than an initial quick demo of base R, I will demonstrate R using RStudio, a GUI for R that is superior in many ways, in my opinion. You can find it here, also free:
http://www.rstudio.com/products/rstudio/download/
I will not assume that you have used R before, and will teach it from scratch. There are also many good online R tutorials – you can find them easily by searching; here are a few to get you started:
http://www.r-tutor.com/r-introduction
http://www.statmethods.net/about/books.html
There is also this free, self-paced online course developed by a colleague of mine:
http://bigdatauniversity.com/bdu-wp/bdu-course/introduction-to-data-analysis-using-r/

Academic Integrity

Obviously, it is an academic offence to use or provide other students with unauthorized aids during quizzes and term tests. Unauthorized aids include but are not limited to: notes, cell phones, another student's paper (direct copying), whispering answers, etc.

Especially note that **it is an academic offence to present someone else's work as your own, or to allow your work to be presented for this purpose**. To repeat: the person who allows her/his work to be copied is equally guilty, and subject to disciplinary action by the university.

Here are some guidelines that apply to the computer assignments. If there is a problem with plagiarism, it will probably happen here, since computer assignments will be handed in.

- It is permitted to copy from me. If your work is very similar to what is presented in lecture, office hours or suggested readings, that is okay. Copy parts of it or use it any other way you like; you are responsible for the results.
- If two students have computer work that is excessively similar to each other, but *not* similar to what was presented in lecture or office hours, that is evidence of cheating. Of course it's easier to detect if the work is also wrong.

- If you allow anyone to have an *electronic* copy of your computer work, for any reason, you are not only guilty of an academic offence, you have also lost your mind.
- *Direct copying of computer code from the internet is prohibited*. You are expected to do the work yourself.
- *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 code from anyone, or allow your code to be copied.
- Suppose you have finished the assignment, and a friend who has not started yet asks you for help. This "friend" is out of line. He or she is expected to give the assignment a serious try before seeking help.
    - But you don't have to be rude. If you want to help, you can start by finding out if the person knows what the computer assignment is asking students to do. You may find that the person has not even read the question yet.
    - Once the meaning of the assignment is clear, you might try walking the person through a set of overheads that is similar to the assignment (there will surely be one). You can ask things like "Do we really need to do this part," and "Well, what does the question say?"
    - It is also okay to compare numerical answers. Questions like "What did you get for beta-hat-4? I got -7.23" are perfectly acceptable.
- But never, ever give an *electronic* copy of your output file to anyone before the quiz. The danger that it will be directly turned in (or transmitted to someone else who will turn it in) is too great. Nobody will believe it was an accident or misunderstanding. The response will be that you *should have known* it might be used as an unauthorized aid. That's in the academic code.
- Comparing output files is acceptable, but **comparing program files is not permitted!** To avoid being charged with an academic offence, do not allow anyone in the class to see your R program file before a computer assignment is due, and do not look at anyone else's. This includes a quick peek at your computer screen. Some people have fantastic memories.
- 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.

Above all, **don't copy, and don't let anyone else copy from you**. You are expected to do the work yourself, and then *perhaps* compare answers after you have done so.

For more detail, 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 The Academic Regulations of the University are outlined in the Code of Behaviour on Academic matters, which can be found in the Arts and Science Calendar or on the web at http://www.governingcouncil.utoronto.ca/policies/behaveac.htm.

Course Schedule

This schedule represents the ***slowest*** we will possibly move through the material.
I certainly hope to move faster, in which case we'll work ahead. Please be prepared.

Lecture no.	L0101 Date	L0201 Date	Hrs	Textbook Reference	Topic	Textbook Practice Problems
1	15-Sep	17-Sep	2	Ch 1	Introduction to course Types of data, SLR model	1.3, 1.5, 1.6, 1.7, 1.8, 1.29, 1.30
2	17-Sep	17-Sep	1	Ch 1	SLR estimation Properties of LS fitted line	1.11, 1.16, 1.20*, 1.33
3	22-Sep	24-Sep	2	Ch 1	Maximum Likelihood Estimation Properties of LS estimators	1.18, 1.21*, 1.24*, 1.36, 1.39a, 1.40, 1.41
4	24-Sep	24-Sep	1		Introduction to R	
5	29-Sep	1-Oct	2	Ch 2	Review of Distribution Theory	
6	1-Oct	1-Oct	1	Ch 2	Review of Distribution Theory	
7	6-Oct	8-Oct	2	Ch 2	Inference for SLR	2.1, 2.2, 2.3, 2.5*, 2.6* (skip e), 2.8 (skip b), 2.10, 2.12, 2.14*, 2.15*, 2.18, 2.51, 2.52
8	8-Oct	8-Oct	1	Ch 2	ANOVA approach to SLR	2.56
9	13-Oct	15-Oct	2	Ch 2	Correlation	2.21, 2.22, 2.24*, 2.25*, 2.35, 2.36, 2.54
10	15-Oct	15-Oct	1	Ch 2	Dummy Variables	**Assignment #1 due**
11	20-Oct	22-Oct	2	Chs 1-2	**Midterm** L0101 (10-12); L0201 (5-7)	
12	22-Oct	22-Oct	1	Ch 3	Diagnostics for X Leverage	
13	27-Oct	29-Oct	2	Ch 3 Ch 10.3-4	Diagnostics for Residuals Influence metrics for X	3.1, 3.2, 3.4 (b, d, e – no Table lookup, f, h), 3.5 (b-f) 10.9efg
14	29-Oct	29-Oct	1	Ch 3	Variable transformations	3.9, 3.18, 3.19, 3.20
15	3-Nov	5-Nov	2	Ch 4	Joint Inference Regression Through Origin	4.1, 4.3, 4.4, 4.8, 4.19, 4.21, 4.24, 4.25

16	5-Nov	5-Nov	1	Ch 4	Predictor Levels Measurement Errors	**Assignment #2 due**
17	10-Nov				**Fall Break – no class**	
18	12-Nov	12-Nov	1	Ch 5	Matrices review	5.1, 5.3
19	17-Nov	12-Nov	2	Ch 5	Matrix approach to SLR	5.4, 5.8, 5.10, 5.12, 5.17, 5.21, 5.23, 5.29, 5.31
20	19-Nov	19-Nov	1	Ch 6	Multiple distinct predictors Polynomial regression	6.2, 6.3
21	24-Nov	19-Nov	2	Ch 6	Indicator variables, GLMs Diagnostics, Inference	6.5 (a-d), 6.6, 6.7 (a), 6.8, 6.18 (b-e), 6.19, 6.20, 6.21, 6.22, 6.25, 6.27
22	26-Nov	26-Nov	1	Ch 7	Extra sums of squares Partial F-tests	7.2, 7.3, 7.7, 7.8
23	1-Dec	26-Dec	2	Ch 7 Ch 10.5 Ch 9	$R^2_{Y1\|2}$, Collinearity VIF Model Selection	7.12, 7.15, 7.20, 7.22, 7.23, 7.27, 7.31 10.15, 10.18
24	3-Dec	3-Dec	1	Ch 8 Ch 10.1	GLMs, Hidden extrapolations Added-variable plots	10.5, 10.8
25	8-Dec	3-Dec	2		Case Study Missing Data	**Assignment #3 due**

For questions marked *, you should use R to obtain your solution. This is good practice for the assignments. You might have to delay the first few * questions until the R lecture.