STA303 / 1002 HF - Methods of Data Analysis II

Wei(Becky) Lin
Spring 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

- Section L0101
  - Tuesday 10:10-12:00 in MC102 (reading week - no classes: Feb. 20-24).
  - Thursday 10:00-11:00 in MC102 (reading week - no classes: Feb. 20-24).
  - Make sure you are able to write the midterm exam, 10-12, Thursday, Mar 2nd.

- Section L0201:
  - Tuesday 15:10-17:00 in SS2117 (reading week - no classes: Feb. 20-24).
  - Thursday 12:00-13:00 in SS2117 (reading week - no classes: Feb. 20-24).
  - Make sure you are able to write the midterm exam, 11-13, Thursday, Mar 2nd.

How to find the classrooms? Please check out
http://www.osm.utoronto.ca/map/f?p=110:1:0::NO::P1_SEARCH:

Important Dates: 2017 Winter
http://www.artsci.utoronto.ca/current/course/timetable/1617_fw/2017_winter_dates

INSTRUCTOR & TA OFFICE HOURS

- Instructor: Wei(Becky) Lin (wei.lin@mail.utoronto.ca). Office: SS6011
- Office hours (an ideal time to discuss questions that you have)
  - Instructor: Wednesday 12:00-13:00 in SS6011 (starts Jan. 15th)
  - Instructor: Thursday 11:10-12:00 in SS2117 (starts Jan. 12th)
- TA office hours will be scheduled before midterm and assignments due dates. Check out announcement on portal.

In general, I am not able to answer questions about the course material, assignments, and tests by e-mail for hundred students (~ 610). Your understanding and cooperation are highly appreciated. Please don’t ask questions about the course material or assignments that are more appropriately discussed in tutorial 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 website or the discussion board on UT portal, if you do not get a response, this may be why. If you believe that issues can be resolved by email, please put STA303: at the start of your subject, as I teach multiple courses in the spring term.

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 messages, 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, practice problems, and announcements are available on

https://portal.utoronto.ca

Please note that we have a Discussion board on Piazza, a TA is assigned to answer questions you have here. If you post your questions there and don’t get response in 4 days, please inform me ASAP. Here is the signup link:


For active participants, 1 point is added to his/her final exam mark for top 60 student askers and 2 points for the top 30 student answers (for a student in both categories, max =2 points).

**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.)


**EVALUATION**

<table>
<thead>
<tr>
<th>Assignment 0</th>
<th>Weight</th>
<th>Date</th>
<th>Time</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2%</td>
<td>Saturday, Jan. 21st</td>
<td>Due: 10pm</td>
<td>Crowdmart (online)</td>
</tr>
<tr>
<td>Assignment 1</td>
<td>8%</td>
<td>Saturday, Feb. 4th</td>
<td>Due: 10pm</td>
<td>Crowdmart (online)</td>
</tr>
<tr>
<td>Midterm</td>
<td>25%</td>
<td>Mar. 02 (L0101), Mar. 02 (L0201)</td>
<td>10:00-12:00</td>
<td>MC102, ...TBA</td>
</tr>
<tr>
<td>Assignment 2</td>
<td>10%</td>
<td>Sunday, Feb. 26th</td>
<td>Due: 10pm</td>
<td>Crowdmart (online)</td>
</tr>
<tr>
<td>Assignment 3</td>
<td>10%</td>
<td>Saturday, Apr. 1st</td>
<td>Due: 10pm</td>
<td>Crowdmart (online)</td>
</tr>
<tr>
<td>Final Exam</td>
<td>45%</td>
<td>Available on Feb. 17th</td>
<td>3-hour exam</td>
<td>TBA</td>
</tr>
</tbody>
</table>

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). The midterm papers are same for both sections, L0101 and L0201. 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 on Thursday, please enrol the course only if you are able to write the midterm.

No late assignments will be accepted without documentation of a valid reason. Late assignment without a valid reason will get penalty of 10% per day off.

Any requests to have marked work re-evaluated must be made within 7 days (one week) as instructed of the date the work or result was returned to the class. The request email must contain a justification for consideration and your clear section information. 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. Regarding how to install R and RStudio, and 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 I found very helpful:


*Introduction to R* by Vecables, Smith and others.

For assignment, you will use Rmarkdown to write your solution (PDF submission is preferred). The first assignment, I will provicue you the template. To learn more about R markdown, refer to

http://rmarkdown.rstudio.com

First time user of R markdown:

https://www.youtube.com/watch?v=QaKCirYkmS8

R Markdown short tutorial with RStudio:

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

**ACADEMIC INTEGRITY**

It is academic dishonesty 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 computer 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. Never look at anyone else’s printouts or show anyone yours before the quiz or exam when they might be handed in.

- 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 we catch you, you will get in big trouble.

- 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](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/behavac.htm](http://www.governingcouncil.utoronto.ca/policies/behavac.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 legitimate to discuss assignment problems with other students in the class or discussion board on portal. 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.

## 1 Course Schedule

The tentative schedule, as of Feb 02nd, 2016, of STA303/1002 course follows. The schedule may change as circumstances necessitate.

<table>
<thead>
<tr>
<th>Week</th>
<th>Dates</th>
<th>Reference</th>
<th>Topics</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>05-Jan</td>
<td>KNN: Ch8</td>
<td>Introduction to course</td>
<td>Try first R markdown le-1.pdf test.pdf -&gt; 3 pdf files</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SJS: CH1</td>
<td>Introduction to R markdown</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10-Jan, 12-Jan</td>
<td>KNN: Ch 8</td>
<td>t-tests</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SJS: Ch 1</td>
<td>1-way ANOVA</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>17-Jan, 19-Jan</td>
<td>SJS: Ch 2</td>
<td>2-way ANOVA</td>
<td>A0 due: 10pm, Jan 21st.</td>
</tr>
<tr>
<td>4</td>
<td>24-Jan, 26-Jan</td>
<td>KNN: Ch 2</td>
<td>ANCOVA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SJS: Ch 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>31-Jan, 02-Feb</td>
<td>KNN: Ch 11.1</td>
<td>Weighted Least Squares (WLS) Regression</td>
<td>A1 due: 10pm, Feb 4th.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SJS: Ch 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>07-Feb, 09-Feb</td>
<td>KNN: Ch 11.2</td>
<td>Ridge regression</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>14-Feb, 16-Feb</td>
<td>KNN: Ch 14.1-14.4</td>
<td>Logistic regression</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SJS: Ch 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>20 to 24-Feb</td>
<td>Reading week, no classes</td>
<td>A2 due: 10pm, Feb 26th</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SJS: Ch 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>07-Mar, 09-Mar</td>
<td>KNN: Ch 14.13</td>
<td>Logistic regression with replicates</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SJS: Ch 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>14-Mar, 16-Mar</td>
<td>KNN: Ch 5</td>
<td>Poisson regression</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SJS: Ch 10</td>
<td>Log-linear model for count data</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>21-Mar, 23-Mar</td>
<td>SJS: Ch 10</td>
<td>Repeated Measures ANOVA, Linear mixed effect model</td>
<td>A3 due: 10pm, Apr 1st</td>
</tr>
<tr>
<td>13</td>
<td>28-Mar, 30-Mar</td>
<td>SJS: Ch 10</td>
<td>Repeated Measures ANOVA, Linear mixed effect model</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>04-Apr</td>
<td>SJS: Ch 10</td>
<td>Linear mixed effect model</td>
<td></td>
</tr>
</tbody>
</table>
STA 304H1 S/1003H S, WINTER 2017
SURVEYS, SAMPLING AND OBSERVATIONAL DATA
Time: M 4-5, AH 100, Th 3-5, NF 003, web-site: on Portal.

Instructor: Dragan Banjevic (dragan.banjevic@utoronto.ca), office BA8139, tel: 946-3939, office hours: Wednesdays, 5-6.

Textbook: Scheaffer, Mendenhall, Ott: Elementary Survey Sampling (Seventh ed.).
Useful but not required: Lohr: Sampling: Design and Analysis.

Marking scheme: First test (20%, February 9), second test 20% (March 20, tentatively) (tests are held in class time), final exam 60% (3h, in exam period, April 10-28). Formula sheet for tests and final will be posted on Portal. There are no make-up tests. With a valid reason (U of T doctor’s note) your mark will be adjusted. If you miss the first test, the second test weight will be adjusted. If you miss the second test, the weight of the final will be adjusted (warning: difficulty increases from the first test to the final; final covers complete course).

Tutorials: There are no tutorials, but you can come for help to Stat. Aid Centre, SS1091, before tests: date and time will be announced. Some extra office hours before the final will be available. Class slides and sample tests and finals will be posted on the web-site. You are required to bring handouts to the class regularly.

Calculation: No statistical software is required. Still, the course includes a lot of numerical calculation. You will need a basic scientific hand-calculator, with statistical functions, and experience in working with it (start using it from the first day). Inability to work with it will not be an excuse. Programmable calculators are not allowed on tests and final exam. Don’t forget this.

Course outline: Almost all of the course material is covered by the textbook. Related to the basic level of the textbook, some theoretical results will be considered in more detail. The following is a tentative schedule for the course:

1. Sampling problems and notions (Ch 2), recommended exercises: 1-7, 28.
2. Basic concepts (Ch 3). Exercises: 2-8, 21.
3. Simple random sampling (Ch 4; 4.6 is not covered), exercises: 1, 2, 14-17, 18a, 22, 21, 23-28, 36, 38, 41, 42.
4. Stratified random sampling (Ch 5; 5.10, 5.11 are not covered), exercises: 1-3, 5-8, 12-17, 24, 26, 27.
5. Ratio, regression, and difference estimation (Ch 6; 6.5 is not covered), exercises: 1, 2, 6, 9, 16, 23, 26, 27.
6. Systematic sampling (Ch 7), exercises: 3, 4, 8, 21, 25, 27.
7. Cluster sampling (Ch 8; 8.8 is not covered), exercises: 2-5, 8, 9, 16, 17, 20, 24, 25, 26, 27.
8. Two-stage cluster sampling (Ch 9), exercises: 2-4, 6, 9, 10, 14-16.
9. Supplemental topics, nonsampling errors (Ch 11.1, 11.2, 11.4, 11.8 are covered), exercises: 1, 13, 14.