SOSC5090: Quantitative Methods to Social Science I
Fall 2023 (tentative)
Mondays 10:30 am-1:20 pm
Rm 5566 (lift 27-28)

Instructor

HAN Li
Contact Information: SOSC Rm. 3390; Tel: 2358-7838; lihan@ust.hk
Office Hours: Mondays 2:30 --3:30 pm (or by appointment)
Course Website: Canvas

Tutorial Instructor

Ms. YIN Yabin
(yyinak@connect.ust.hk)
Office Hours: Tuesdays 2-3 pm @ Rm3001

Course Description

The purpose of this course is to introduce basic statistical concepts and applications that are heavily used in (quantitative) social science research. The course serves as a foundation for anyone who is interested in empirical research and also as a prerequisite to taking more advanced methodology courses (such as SOSC5340 and above).

After completing this course, you will be able to:

- Understand basic statistical concepts and methods used in social science research
- Use basic function of R and Stata to analyze real data
- Given a research question, use relevant data to test hypotheses, conduct statistical inference, and interpret regression results
- Clearly present your research work, in both oral and written formats

Required Text

The required textbook for this course is *Quantitative Social Science: An Introduction* by Kosuke Imai, which is available at the bookstore and also in the library. Supplementary readings for the course come from several chapters and appendices (see the list of chapters in the course outline) in *Introductory Econometrics: A Modern Approach*, 4e (2009) by Jeffrey Wooldridge (reserved in the library too). Other readings (such as research papers) will be uploaded to the course website when needed.

Software

You are required to use R and/or Stata to do statistical work in this course. While using Stata is acceptable for some exercises and final paper, we emphasize how to use R to answer quantitative social science questions in lectures.

R is freely available for download and runs on Macintosh, Windows, and Linux computers. Students are strongly encouraged to use Rstudio, another freely available software package that has numerous features to make data analysis easier. Install package ***swirl*** and find some course exercises there. Some of questions in problem sets are based on the exercises.

Additional resources:

- Download R: https://www.r-project.org/
- Rstudio: https://www.rstudio.com/
- An Introduction to R: https://cran.r-project.org/doc/manuals/R-intro.pdf
- UCLA IDRE: https://stats.idre.ucla.edu/r/
- Princeton R Tutorial: http://data.princeton.edu/R
- A Quick Introduction to R (for Stata Users): http://rslblissett.com/wp-content/uploads/2016/09/RTutorial_160930.pdf
- You can combine the Swirl exercises for this course with the exercises for another course named R programing.
- R for beginners. https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

Stata14 is installed in all PCs at social science computing lab (Rm. 3001). You can also have access to STATA network version through Virtual Barn Desktop provided by HKUST. For many of you who have never used Stata before, the TA will help you get familiar with Stata in tutorials. It is also easy to train yourself given the rich self-learning resources available. Here are some recommended resources:

- Hamilton, Lawrence C. (2006). *Statistics with STATA*. Belmont, CA: Duxbury Press. (A copy is available for loan at TA's office.)
- UCLA Stata Portal (an extensive resource that leads you to many useful links): http://statcomp.ats.ucla.edu/stata http://www.ats.ucla.edu/stat/stata/sk (Starter Kit section for new users)
- Princeton Stata Tutorial: http://data.princeton.edu/stata/
- UNC Carolina Population Center: http://www.cpc.unc.edu/services/computer/presentations/statatutorial/

Course Requirements and Grading

Your final grade in the course will be based on your performance on the following:

- Class participation (lectures & tutorials) 10%
- Assignments 10%
- Take-home exam (Dec 1) 30%
- An independent term paper/replication paper
  - ✓ Final paper presentation 20%
  - ✓ Final paper write-up 30%

We have six assignments through the semester in total and only count 5 with highest grades. Each accounts for 2% of the final grade. The main purpose of the assignments is to make you familiar with the statistical concepts and comfortable with processing the data.

For the independent project, you will find a published article that interests you, preferably but not necessarily related to one of the topics covered during the semester. I will suggest some articles. Your basic task is to replicate that article. Replication means obtaining the original data

the author(s) used (or essentially similar data, but **NOT** the data extract accompanying the article that can be downloaded from some journals' websites), reconstructing the sample(s) the author(s) used for statistical/regression analysis, repeating the main results the author(s) obtained, performing some robustness/sensitivity tests the author(s) conducted (for example, those tests only mentioned in the article text but not shown in the tables -- typically for space concerns), and (encouraged but not required) even doing some updating or extension work if you think the article has limitations or room for improvement.

At first, this project may sound like a mechanic exercise, but you will find that it is challenging and something from which you can learn much. You will learn about the problems encountered in empirical research and see that published results are often not nearly as neat and clean as they seem. Indeed, replication of existing findings is where you should usually begin any empirical project.

Your grade of the replication project will be based on an oral presentation in the class (Nov 27) and a written paper due on **20 December**. The paper should be no more than 30 (double-spaced) pages and should at least include an introduction, a brief literature review summarizing the original article's (or/and your replication's) contribution, a description of data and sample construction, a comparison of original results and your replication (including relevant discussions), and references, tables, and figures (if applicable). Only a printed copy of your paper is accepted. For my records, you are also required to submit your data and R code file or Stata do-files along with your paper, although they will not be graded. More guidelines on how to present and write up your work will be provided during the semester.

I expect each of you to meet with me to discuss about your projects, and I will be happy to offer advice and help at any stage of the process. To make sure you will make a good progress and not leave everything till the last minute, please notify me (by email is sufficient except for the final paper) of the following stages by their dates (none of the stages before the final paper will be graded).

- 18 September: 1 paragraph (article to replicate, data to use, and specific aims)
- 16 October: descriptive statistics; preliminary results
- 11 November: replication results for presentation; an outline of the paper
- 27 November: presentation
- 14 December: final paper (with data and do-files or R files) due

Problem set collaboration policy

Problem sets for this course present opportunities for students to discuss questions and collaborate to find a solution together. At the same time, as with any class that includes analytical exercises and computer programming, there is a clear distinction between permissible collaboration and unacceptable plagiarism. This course will follow a modified version of the guidelines of the university. Please take this guideline seriously. In the past, plagiarism cases typically resulted in at least the failure of a course.

Programming necessitates that you reach your own understanding of the problem and discover a path to its solution. During this time, discussions with other people (whether via the Internet or in person) are permitted and encouraged. However, when the time comes to write code that solves the problem, such discussions (except with course staff members) are no longer

appropriate: the code must be your own work. Do not, **under any circumstances**, copy another person's code. Incorporating someone else's code into your program in any form is a violation of academic regulations. Abetting plagiarism or unauthorized collaboration by sharing your code is also prohibited. Sharing code in digital form is an especially egregious violation: do not e-mail your code to anyone. Novices often have the misconception that copying and mechanically transforming a program (by rearranging independent code, renaming variables, or similar operations) makes it something different. Actually, identifying plagiarized source code is easier than you might think. For example, there exists computer software that can detect plagiarism. If you have any questions about these matters, please consult the course instructor.

For ALL TYPES of assessments (including problem sets, exams, papers, and so on), ALWAYS WRITE SEPARATELY!!!!!

Course Outline and Tentative Schedule

Note: **QSS** stands for *Quantitative Social Sciences: An Introduction* and **JW** stands for *Introductory Econometrics: A Modern Approach*. **Ch** stands for chapters and **Ap** stands for Appendix.

1. Week 1: Motivation, Overview and a brief introduction to R (QSS Ch1, JW Ap A)
2. Week 2-3: Causality (QSS Ch2)
    a. An example of randomized trials
    b. Observational studies
3. Week 4-5: Measurement and Research Design (QSS Ch3, JW Ap B)
    a. Survey sampling and its potential biases.
        e.g., Measuring public opinions through sample surveys
    b. Clustering
4. Week 6-7: Prediction (QSS Ch 4, JW Ch2-5)
    - Prediction and loop
    - Linear regression models
5. (If time permits) Discovery of patterns from data of various types (QSS Ch5)
    a. textual data
    b. network data
    c. geo-spatial data
6. Week 8-9: Probability theory, random variables, and probability distribution (QSS Ch 6, JW Ap B)

a. Probability and conditional probability

b. Random variables and their distributions, large number theorems

7. Week 10-11: Uncertainty (QSS Ch 7, JW Ap C)

a. Estimation

b. Hypothesis testing and confidence intervals

b. Regression with uncertainty

8. Last week: student presentations & Take-home exam