SOSC 2400 **Quantitative Data Analysis for Social Research II**

Fall 2022
Wednesday & Friday, 3:00-4:20pm
Room 2465 (Lift 25-26)

Instructor: Dr. WANG Hongbo (hbwang@ust.hk)
Office: Academic Building, Room 2372 (Ext. 7804)
Office Hours: By appointment

TA: CHEN Ziyang (zchenes@connect.ust.hk)
Office: Academic Building, Room 3001
Office Hours: By appointment
LIU Siqin (sliuca@connect.ust.hk)
Office: Academic Building, Room 3001
Office Hours: By appointment

Course Description and Objectives:
This course mainly covers implementation of linear regression from a social scientific perspective. The focus is on the specification of models including choice of variables, incorporation of different types of effect of X on Y, and interpretation of coefficients. Note that the course will not treat statistical inference explicitly, leaving the latter almost entirely to a formal statistics course.

Organization:
The class meets twice a week on *Wednesday* and *Friday*, respectively. The lecture will be held on *Friday*, while *Wednesday* usually is reserved for a computing session following the lecture (See "Schedule" below).

All course materials will be distributed through Canvas. Note that they should be used *exclusively* for the purpose of this course.

Students are encouraged to form groups of up to 5 members to collaborate on the course project (See below for details). Although the division of work is decided entirely by the group themselves, every member is expected to make adequate contribution to the project. *Free-riding is considered a cheating behavior and will be panelized as such.*

Computing:
We will mainly use R for computing.

Prerequisite:
SOSC 1100

References:

Baumer, Benjamin S., Daniel T. Kaplan, and Nicholas J. Horton. 2017. *Modern Data Science with R*. Chapman and Hall/CRC. [**BKH**]

Imi, Kosuke. 2018. *Quantitative Social Science: An Introduction*. Princeton University Press. [**I**]

Navarro, Danielle. Learning statistics with R: A tutorial for psychology students and other beginners (Version 0.6). [**N**]

Treiman, Donald J., 2009. *Quantitative Data Analysis: Doing Social Research to Test Ideas*. Jossey-Bass. [**T**]

Assessment:

Your grade will be determined as follows:

(1) Attendance: 20%

Attendance is crucial and required for all students and all classes. One point will be deducted from your attendance score for each class missed *without any legitimate reason*.

(2) Participation: 10%

Your class participation will be evaluated regarding in-class exercises, discussion, and so on. The participation score is determined using a four-level scheme, i.e. excellent (=10), satisfactory (=8), unsatisfactory (=5), and completely fail (=0). Note that *an attendance score less than 10 automatically leads to "fail" on participation*.

(3) Group project: 70% (presentation, 20%; written report, 50%)

Under the supervision of the instructor, each group will choose a topic of their own, locate appropriate data sources, carry out data analysis, present the findings, and, finally, submit a written paper. Detailed guidelines for the project can be found in Appendix.

Each group should keep a diary of their work on the project, which describes all related activities, including exploration of literature and data, downloading and processing of data, and analyzing data. It should indicate all efforts group members have made on the project, including on data processing and exploration not reflected in the final report.

Every group will need to submit a brief report on the progress of the project in the middle of the semester (See Schedule below).

Schedule (*Subject to modification*)

Academic Calendar Week	Topics		Important Events
Week 1: Friday (9/2) Wednesday (9/7) Friday (9/9)	**Course overview** **[R]** R computing: A review **[L]** – Introduction to SLR – Guidelines for Project		**Project group finalized**
Week 2: Wednesday Friday	**[R]** *R Practice* **[L]** Choice of X		
Week 3: Wednesday Friday	**[R]** `ggplot2` and `stargazer` **[L]** Coefficient and Goodness-of-fit		**Initial proposal due**
Week 4: Wednesday Friday	**[R]** Descriptive Analysis **[L]** Marginal effect		
Week 5: Wednesday Friday	**[R]** Fitting regression using `lm()` **[L]** Log transformation		
Week 6: Wednesday Friday	**[R]** SLR with logged variables **[L]** Curvilinear Effect		
Week 7: Wednesday Friday	**[R]** Incorporating polynomials in regression **[L]** Categorical X		
Week 8: Wednesday (10/26) Friday	**[R]** Factor Variable **[L]** Interaction Effect		**Mid-term report due**
Week 9: Wednesday Friday	**[R]** SLR by Group **[L]** Motivation for MLR		
Week 10: Wednesday Friday (11/11)	**[R]** *R Practice* **[L]** *Project Proposal (Presentation)*		**Final proposal due**
Week 11: Wednesday Friday	**[R]** *Q & A* **[L]** Course Review		
Week 12: Wednesday (11/23) Friday	**[R] Project Presentation** **[L] Project Presentation**		**Presentation PPT due**
December 6			**Final report due**

[L] Lecture
[R] R computing

Appendix:

Guidelines for Project

- The project should address an empirical question of social scientific relevance by applying simple linear regressions (SLR) to real-world data. If you need advice on how to conceive insightful research questions, Firebaugh (2008) can be a good reference. The empirical question needs to involve one quantitative dependent variable (Y) and three independent variables (X's). The project should choose three X's that affect Y in different ways, respectively. Specifically, X's need to contain at least one continuous variable and one categorical one. The effect of each X on Y will be examined using SLR as properly specified. Each group should also locate suitable dataset for the project.

- An *initial* project proposal (font 12, double spaced, and up to 2 pages) is due in **Week 3**. The initial proposal should clearly specify Y and three X's at the minimum. The choice of variables needs to be justified on the basis of theories, previous studies, or empirical observation. You should also have located suitable data sources. Note that the data need to be carefully checked for sample size, availability of variables, missingness, and so on.
  However, you may leave undecided the specification of the relationship between Y and each X at this stage. Instead, you are allowed to continue to vet and improve the proposal as the course goes. Thus, the proposal also needs to include a work schedule outlining anticipated progress on a weekly basis.

- The *final* proposal, due in **Week 10**, needs to clearly specify how each X affect Y in a regression. Is the effect linear, curvilinear, or what? *You are not allowed to make any major change after submitting the final proposal*.

- Every group will present main findings from their project in the last week.

- The final report (font 12, double spaced, up to 30 pages), due on **December 6**, should include the following parts:

1. Title page with title, authors, date, and division of work §1 page
2. Motivation and empirical questions §1 page
3. [*Optional*] Literature review summarizing research questions and key findings from one relevant previous study §1 pages
4. Testable hypotheses regarding the relationship between Y and each X §1-2 pages
5. Description of data, variables/measurement, and data processing §1-2 pages
6. Descriptive statistics (univariate distributions and correlational analysis), shown in publishable tables and graphs, with adequate interpretation §2-3 pages
7. SLR model specifications regarding the relationship between Y and each X §1-2 pages
8. SLR results, shown in publishable tables and graphs, with adequate interpretation §10-15 pages

9. Summary of key findings followed by discussions §1-2 pages
10. References §1 page
11. R scripts and supplementary files

References

Firebaugh, Glenn, 2008. *Seven Rules for Social Research. Princeton*. New Jersey: Princeton University Press. [Chapter 1]