ECON3284: Introduction to Causal Inference and Statistical Learning

Instructor: Yanhui Wu
Email: yanhuiwu@hku.hk
Office: K.K. Leung 931
Phone: 29178508
Consultation time: 4-5pm (Monday and Thursday) or by appointment

Course description: This course introduces students to fundamental ideas and important techniques from modern causal inference and statistical learning. Combining basic statistical theory, scientific principles of research design, and hands-on experience with real data, this course will provide students a solid basis for being good consumers and practitioners of empirical research in economics and other quantitative social sciences. The course will be divided into two parts. The first part (3/4 of the classes) will concentrate on causal inference, and the second part (1/4 of the classes) on statistical learning, covering elementary topics in statistical prediction and machine learning. Applications will be drawn mostly from economics and at times from other disciplines to broaden students’ horizon.

The primary focus of this course is on application instead of methodological rigor. Hence, the use of mathematics will be limited to elementary algebra and probability. However, students are expected to have taken introductory courses in both statistics (econometrics) and microeconomics. Because of the emphasis on hands-on data experience, students are expected not to be scared by data and programming. Previous experience with statistical software and knowledge about computer programming is an advantage but not required. Homework assignments are designed to familiarize students with the necessary programing language.

We will use the statistical package R via a front-end called RStudio. This is a standard statistical software used by social scientists because of its powerful function. Both R and RStudio are free and open source. It is also fine if a student prefers Stata, another software that is popular among economists. But using Stata requires a license.

Textbooks: The following two books will be heavily used in the course. The second one (ISLR) is important for homework assignment.

Assessment:

Since we put emphasis on hands-on experience, learning by doing and class participation will be key to success in this course. There will be six problem sets throughout this course. Except for the first one, all the other problems, as homework assignment, will be graded. These five problem sets will account for 40% of your score, each 8%. Class participation will account for 10%.

At the end of the semester, there will be a final exam and the term project, each account for 25% of your score. The final exam will cover the theoretical concepts and applications discussed in class. The term project is mostly coding exercises in which students (formed in groups) are required to write R codes to replicate the empirical results in some existing studies. Details about project assignment and group formation will be announced in the middle of the course.

Course Policy:

1. This is an active learning course, and attendance and participation are extremely important. Please observe appropriate classroom etiquette and be considerate to others. In particular, laptop use should be limited to course-related activities, and cell phones are not allowed in class.

2. Students are encouraged to work together in groups to solve the problem sets. However, each student must turn in his or her own homework. Copying another student's answers is not permitted even with consent. All assignments including the term project report must be typewritten.

3. Plagiarism and cheating in exams are serious academic offenses.
Course Material and Class Schedule

The course will be divided in two types of class: “story-telling” (Lecture) and “game-playing” (Lab). Typically, the Monday long class (12:30-14:20 pm) is devoted to lectures, and the Thursday short class (12:30-13:20pm) is devoted to labs. The lecture class will focus on empirical methods and their applications based on textbooks and original papers. For the original papers, you are not expected to know every detail. Our focus will be on methodology instead of economic substance. In the lab class, we will discuss problem sets, particularly those involving computer programing. It should be noted that we are not expected to do the coding and debugging work in class. Rather, we will talk about ideas, results, and common problems. You are expected to do the coding at home and submit your answer together with the codes before the due date.

The course consists of eight topics listed below. The corresponding material of each topic is classified into: Lecture (material that will be covered intensively during the lecture class) and Lab (material that will be discussed during the lab class).

**Topic 1. Introduction and review of basic statistical concepts**

(Lecture) AP-MM, chapter 1, appendix: Mastering Inference

(Lecture) AP-MM, chapter 2, appendix: Regression Theory

(Lab) ISLR, Chapter 2.3: Introduction to R

(Lab) ISLR, Chapter 3.6: Linear Regression

(Lab) Problem set 1: Basic econometrics using R

**Topic 2. Randomized experiments (RCT: randomized controlled trial)**

(Lecture) AP-MM, chapters 1-2


(Lab) Problem set 2: Randomization and causal inference

**Topic 3. Instrument variables (IV)**

(Lecture) AP-MM: chapter 3


(Lab) Problem set 3: IV

**Topic 4. Regression discontinuity (RD)**

(Lecture) AP-MM: chapter 4


(Lab) Problem set 4: RD

**Topic 5. Panel data and Difference-in-differences (DID)**

(Lecture) AP-MM: chapter 5


(Lab) Problem set 5: DID

**Topic 6. Statistical prediction and basics of machine learning (ML)**

(Lecture) ISLR, Chapters 2 and 3

(Lab) Problem set 6: Basic ML

**Topic 7. Classification**

(Lecture) ISLR, Chapter 4

**Topic 8. Clustering**

(Lecture) ISLR, Chapter 10
### Tentative Class Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Monday (12:30-2:20)</th>
<th>Thursday (2:30 -3:20)</th>
<th>Homework due</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>Lecture 1: introduction</td>
<td>Lecture 1: review of statistics</td>
<td></td>
</tr>
<tr>
<td>Week 2</td>
<td>Lecture 1: review of statistics</td>
<td>Lecture 2: RCT, method</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lecture 2: RCT, method</td>
<td>Lab 1: Basic R</td>
<td></td>
</tr>
<tr>
<td>Week 3</td>
<td>Lecture 2: RCT, application</td>
<td>Lab 2: RCT, Part I</td>
<td>PS 2 Part I</td>
</tr>
<tr>
<td>Week 4</td>
<td>Lecture 3: IV, method</td>
<td>Lab 2: RCT, Part II</td>
<td>PS 2 Part II</td>
</tr>
<tr>
<td>Week 5</td>
<td>Lecture 3: IV, application</td>
<td>Lab 2: RCT</td>
<td>PS 2 Part I</td>
</tr>
<tr>
<td>Week 6</td>
<td>Lecture 4: RD</td>
<td>Lab 3: IV, Part I</td>
<td>PS 3 Part I</td>
</tr>
<tr>
<td>Week 7</td>
<td>Lecture 5: DID, method</td>
<td>Lab 3: IV, Part II</td>
<td>PS 3 Part II</td>
</tr>
<tr>
<td>Week 8</td>
<td>Lecture 5: DID, application</td>
<td>Lab 4: RD</td>
<td>PS 4</td>
</tr>
<tr>
<td>Week 9</td>
<td>Lecture 6: Basic ML</td>
<td>Lab 5: DID, Part I</td>
<td>PS 5 Part I</td>
</tr>
<tr>
<td>Week 10</td>
<td>Lecture 6: Basic ML</td>
<td>Lab 5: DID, Part II</td>
<td>PS 5 Part II</td>
</tr>
<tr>
<td>Week 11</td>
<td>Lecture 7: Classification</td>
<td>Lab 6: Basic ML, Part I</td>
<td>PS 6 Part I</td>
</tr>
<tr>
<td>Week 12</td>
<td>Lecture 8: Clustering</td>
<td>Lab 6: Basic ML, Part II</td>
<td>PS 6 Part II</td>
</tr>
</tbody>
</table>

**Notes:** Lecture 1 means the lecture session of topic 1 and Lab 1 means the lab session of topic 1, so on and so forth. The first lab session will start in Week 3. Each to-be-graded homework (PS 2 to 6) is due before the corresponding Lab session. Homework turned in after the corresponding lab session will not be graded.