POL 572: Quantitative Analysis II

Spring 2010

Kosuke Imai
Department of Politics, Princeton University

1 Contact Information

Office: Corwin Hall 036
Office Phone: 258–6601
Email: kimai@Princeton.Edu
URL: http://imai.princeton.edu

2 Logistics

- Lectures: Mondays and Wednesdays 9:00am –10:20am, 101 Sherrerd Hall
- Precepts (taught by Insong Kim insong@princeton.edu):
  Tuesdays 5:00pm –6:30pm, Corwin Hall 127
- Kosuke's Office hours: Stop by anytime or make an appointment
- Insong's Office hours: Wednesdays 4:00pm –6:00pm, Corwin Hall 023

3 Questions about the Course Materials

In addition to precepts and office hours, please use the *Discussion Board* at Blackboard when asking questions about lectures, problem sets, and other course materials. This allows all students to benefit from the discussion and help each other understand the materials. Insong will be primarily responsible for handling questions about precepts and problem sets, while I will primarily responsible for answering the questions about the lectures and other course materials. But, everyone is also encouraged to participate in discussions and answer any questions that are posted.

4 Course Description

This course is the first course in applied statistical methods for social scientists. We begin by studying the fundamental principles of statistical inference. Students will then learn a variety of basic *cross-section* regression models (as time permits!) including linear regression model, structural equation and instrumental variables models, discrete choice models, and models for missing data and sample selection. Unlike traditional courses on applied regression modeling, I will emphasize the connections between these methods and causal inference, which is the primary goal of social science research.

5 Prerequisites

There are three prerequisites for this course.

1. Mathematics covered in POL 502: Basic real analysis, calculus, and linear algebra
2. Probability covered in POL 571: See DeGroot and Schervish (2002) Chapters 1–5.
3. Statistical computing covered in the statistical software workshop held at the beginning of the academic year. The workshop materials are posted at Blackboard.

6 Course Requirements

The final grades are based on the following items:

- **Problem sets** (50%): Several (almost weekly) problem sets will be given throughout the semester. Each problem set will contain both analytical and data analysis questions and equally contribute to the final grade. The following instructions will apply to all problem sets unless otherwise noted.
  - Neither late submission nor electronic submission will be accepted.
  - Although you are allowed to discuss the problem sets with others, you should not copy someone elses answers or computer code. In particular, sharing a paper or electronic copy of your code and answers with other students is strictly prohibited.
  - For analytical questions, you should include your intermediate steps and sufficiently detailed comments. For data analysis questions, include annotated code as part of your answers. All results should be presented so that they can be easily understood. For example, all graphs should have proper axis labels.
- **Midterm exam** (25%): A three-hour closed-book inclass exam given immediately after the spring break. This covers the first half of the course materials.
- **Final exam** (25%): A three-hour closed-book inclass exam given during the exam period. This covers the second half of the course materials.

7 How to Get Most out of this Course

To get most out of this course, the most important thing is to keep up with the new materials that will be introduced every week. Do not leave any questions you may have about new concepts and equations unanswered. For this purpose, I take an interactive lecture style whenever possible. Problem sets are also designed to help you achieve this goal, but you will not be able to solve them efficiently unless you understand the lectures. Use the assigned and supplementary readings to assist your understanding of the lecture materials. In the past, some students find it useful to form a study group and go over the lecture materials as well as the problem sets. Also, use office hours, precepts, and Blackboard if you are unclear about any part of the course materials. Because the materials in the later part of the course (and Quant III) build upon those covered earlier in the semester, falling behind will mean that you will be lost for the rest of the quantitative methods sequence.

Finally, please start the problem sets as soon as you receive them. You cannot learn statistics without doing. Both analytical and data analysis questions constitute integral parts of the course,

and many of you will find them challenging. Data analysis questions often involve the replication and extension of published articles in top journals, and analytical questions test whether you understand the mechanics of the methods covered in the course. Quant II is a "statistics boot camp" which prepares you for Quant III and IV where you will begin to use quantitative methods and conduct independent research. So, the course will be a rewarding and yet painful experience, but by the time you finish it you notice you have built up your "statistics muscle" quite a bit. Good luck!

8 Statistical Computing

A major emphasis of this course is to have students learn how to better present and communicate the results of their statistical analysis in a manner that can be easily understood by the general audience who has little statistical training. To achieve this goal, we use a statistical computing environment, called R. R is available for any platform and without charge at http://www.r-project.org/. In a recent *New York Times* article ("Data Analysts Captivated by R's Power", January 6, 2009), R is described as software that "allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems." If you prefer, you can use other software for parts or all of the problem sets, but no support will be provided by either me or the preceptor.

In the past, I noticed that some students ended up spending an unnecessarily large amount of time debugging their R code for problem sets simply because they do not know how to write a code which is easy for people (including themselves!) to understand. For those of you who have little prior programming experience (and more experienced programmers with bad coding habits), please follow the Google's R Style Guide available at the following URL:

http://google-styleguide.googlecode.com/svn/trunk/google-r-style.html

Also, using an appropriate text editor makes it easier for you to maintain a good programing practice and avoid unnecessary coding mistakes. I recommend the use of Aquamacs for Mac users and WinEdt (together with `R-WinEdt` package) for Windows users. These text editors have useful functionalities such as syntax highlighting and R command recognition.

9 Books

There is no single textbook for this course. However, you may find the following books (listed below in the alphabetical order) useful and some of them are used for this course. They are also available for purchase at the Labyrinth bookstore and on reserve at the library.

10 Course Outline

Each topic is followed by the list of required readings, which will be made available through Blackboard. In addition to these and my lecture slides, I will provide some optional readings and my own lecture notes throughout the semester. All of the readings will be available through either the library or electronic reserve system. As you can see, the list of topics is quite ambitious. My current plan is to spend three weeks on each topic but the plan is subject to change, depending on how students are keeping up with the course.

10.1 Basic Principles of Statistical Inference

1. Descriptive, Predictive, and Causal inference

   Freedman (2009) Chapter 1.

   Rosenbaum (2009) Chapter 1.

   Morgan and Winship (2007) Chapter 2.

2. Identification, Estimation, and Confidence Interval

   Manski (2007) Introduction and Chapter 7.

   DeGroot and Schervish (2002) Sections 7.1–7.5.

3. Hypothesis Testing

   Rosenbaum (2009) Chapter 2 (Skip Sections 2.4.5–2.5).

   DeGroot and Schervish (2002) Sections 8.1, 8.5–8.7, 9.3–9.5

4. Problem Sets: Sample surveys, Randomized experiments

10.2 Linear Regression

1. Simple Regression

   Freedman (2009) Chapter 2.

   DeGroot and Schervish (2002) Sections 10.1–10.3.

   Angrist and Pischke, Section 6.1.

2. Multiple Regression

   Freedman (2009) Chapter 3.

   (Easier) Freedman (2009) Chapters 4 and 5; or (Harder) Hayashi (2000) Chapters 1 and 2.

3. Problem Sets: Sharp regression discontinuity design, Ecological inference

10.3 Structural Equation Modeling

1. Instrumental Variables

   (Easier) Angrist and Pischke Chapters 4 and 6; or (Harder)

2. Direct and Indirect Effects

3. Problem Sets: Fuzzy regression discontinuity design, Causal mediation analysis

10.4 Maximum Likelihood and Regression Models

1. Likelihood Theory

   (Shorter) Freedman (2009) Section 7.1; or (Longer) King (1998) Chapter 4.

2. Bootstrap and Monte Carlo Approximation

   Freedman (2009) Chapter 8.

3. Discrete Choice Models

   (Shorter) Freedman (2009) Sections 7.2–7.3; or (Longer) King (1998) Sections 5.1–5.4.

4. Missing Data and Sample Selection Models

5. Problem Sets: Retrospective sampling design, Multiple imputation