POL 573: Quantitative Analysis III

Fall 2009

Kosuke Imai

This course is the second course in applied statistical methods for social scientists. Building on the materials we covered in POL 572 or its equivalent (i.e., linear regression, structural equation modeling, instrumental variables, maximum likelihood estimation, discrete choice models), students will learn a variety of statistical methods including models for longitudinal data and survival data. 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.

1 Contact Information

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

2 Logistics

- Lectures: Mondays (101 Sherrerd Hall) and Thursdays (023 Robertson Hall) 10:30–11:50am
- Precepts (taught by Teppei Yamamoto tyamamot@Princeton.EDU): Tuesdays 4:30–6:00pm
- Kosuke's Office hours: Stop by anytime or make an appointment by email
- Teppei's Office hours: Wednesdays 4:30–6:30pm (126 Corwin)

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. Teppei will be primarily responsible for handling questions about precepts and problem sets, while I will answer the questions about the lectures and other course materials. But, students are also encouraged to participate in discussions and answer any questions that are posted.

4 Prerequisites

There are three prerequisites for this course.

1. Mathematics at the level of the math camp and POL 502

2. Probability and statistics at the level of POL 571 and POL 572

3. Statistical computing and programming at the level of the statistical software workshop

5 Course Requirements

Your final grade is based on the problem sets and the final project:

- **Problem sets** (40%): There are four problem sets. Although you are allowed to discuss the problem sets with others, you should not copy someone else's computer code or answers. In particular, sharing a paper or electronic copy of your code and answers with other students is strictly prohibited.
- **Final project** (60%): The final project is due 4pm, January 15. Neither electronic nor late submission is allowed. The detailed instructions are given below and will be discussed in the first lecture.

6 Computing

In this course, we support 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, R is described as

> a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca [...] whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it. [...] "The great beauty of R is that you can modify it to do all sorts of things," said Hal Varian, chief economist at Google. "And you have a lot of prepackaged stuff that's already available, so you're standing on the shoulders of giants."

According to the article, R is also software that "allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems."

We choose R for its flexibility and power. However, students may use other statistical software such as STATA for the problem sets and the final project, but at their own risk; that is, we will not be able to answer your software-related questions. Of course, there will be no penalty for using different statistical software. What matters is the analysis you present rather than the software you use.

7 Final Project

The final project is the most important requirement of this course. The goal is to conduct a project that can be *eventually* developed into a high-quality published paper. I encourage you to continue working on your project after this course, and I am happy to continue to provide guidance along the way. You may also consider applying for the summer political methodology meeting using the paper that comes out of your final project. In the past, papers based on final projects for this course appeared in refereed journals and won a graduate student poster award at the political methodology summer meeting. Here are a few important things to note when conducting the final project for this course.

- **Collaboration** with another student in this class is strongly encouraged for at least two reasons. First, collaboration usually leads to a better project given the limited amount of time and intellectual capacity each student has; the sum is typically better than any of its parts and can be even better than the sum of its parts. Second, collaboration will tend to make the project much more fun. You will learn a great deal from each other and will also get to know your collaborator very well. Please contact me if you are considering a collaborative project with someone who is not taking this class.

- **Simultaneous submission to another class** is allowed provided that you obtain a written permission from both instructors in advance. This means that your project must be appropriate for both classes.

- **Two alternative approaches to the project**. First, you may start with an original research question, develop a hypothesis, collect a data set, and conduct statistical analysis of the data in order to test the hypothesis. That is, you conduct a research project from scratch. The second approach is to start with the replication of the results in a published article and improve or extend the original analysis. If you decide to take this approach, you need to contact the author(s) for the replication information as soon as possible (unless it is already available online) so that you will be able to get started on the analysis early in the semester.

- **Proposal** is due in class on December 3. The proposal should contain *at the very minimum* the following components of the final project; (1) the brief statement of a question to be addressed and a hypothesis to be tested, (2) the detailed description of the data set to be analyzed, (3) the detailed results of a preliminary analysis, and (4) the detailed tentative plan of the final analysis. For those of you who have made substantial progress, the "proposal" may take the form of a draft paper.

- **Writing style** is very important for the success of any scientific paper. Be accurate and concise. Pay special attention to tables and figures, which form the core of an empirical paper. Here is what I do when I draft a paper. Start with tables, figures (and their detailed captions), and a rough description of the data and the methods you use (that is, do not start writing until you have them!). Then, write the abstract followed by the introduction. Make sure that you clearly state your contributions in the abstract and at the beginning of the introduction. Once these are done, then it is easy to write the rest of the paper by further elaborating each point you make in the introduction. The structure of the paper should be top-down (i.e., don't keep readers guessing what you are going to say next), and you should carefully write the first and last sentences of each section and paragraph. Take advantage of sectioning to write a paper with a clear and logical structure. There are many books and articles explaining how to write a good scientific paper, but here is one written by a political scientist.

  Gary King. Publication, publication. *PS: Political Science and Politics*, XXXIX(1):119–125, January 2006.

  Finally, I recommend that you consider using LaTeX (together with BibTeX for bibliography).

- **Evaluation** of the final project will be based on the following criteria. For a purely methodological paper, the originality and significance of the proposed method will form the basis of evaluation. The evaluation of a more applied paper depends on how a statistical method is effectively used to answer the substantive question set forth by its author(s).

8 Textbooks

There is no single textbook for this course. Instead, the course materials consist of lecture notes, lecture slides, and assigned readings. In addition, however, you may find the relevant parts of the following textbooks useful. Some of these books are reserved at the Firestone library.

1. Political Methodology

   Gary King. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. University of Michigan Press, Ann Arbor, 1998.

2. Probability and Statistics

   Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics*. Addison Wesley, Boston, 3rd edition, 2002.

   Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, 2005.

3. Econometrics

   Fumio Hayashi. *Econometrics*. Princeton University Press, Princeton, 2000.

   Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, MA, 2002.

4. Regression Modeling

   Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambrdige University Press, Cambridge, 2007.

5. Causal Inference and Research Design

   Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, New York, 2007.

   Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, 2009.

6. R

   John Fox. *An R and S-plus Companion to Applied Regression*. Sage Publications, Thousand Oaks, CA, 2002.

9 Tentative Course Outline

We will cover the following topics in the order they are listed (and as time permits!).

1. Regression Modeling for Cross-section Data (Continued from POL 572)
   (a) Event count models
   (b) Generalized linear models
2. Survival Data Analysis
   (a) Basic concepts
   (b) Parametric regression models
   (c) Cox proportional-hazard model
   (d) Competing risks models
3. Causal Inference with Cross-section Data
   (a) Matching methods
   (b) Weighting methods
4. Regression Modeling for Longitudinal Data
   (a) Linear mixed effects models
   (b) Generalized linear mixed effects models
   (c) Multilevel, Hierarchical models
5. Causal Inference for Longitudinal Data
   (a) Difference-in-difference models
   (b) Matching methods
   (c) Weighting methods
6. Statistical Analysis with Missing Data
   (a) Basic concepts and assumptions
   (b) Multiple imputation