POL 572: Quantitative Analysis II

Spring 2012

Kosuke Imai Carlos Velasco Rivera

Department of Politics
Princeton University

1 Contact Information

	Kosuke Imai (Instructor)	Carlos Velasco Rivera (Preceptor)
Office:	Corwin Hall 036	Corwin Hall 026
Phone:	258–6601	937–5558
Email:	kimai@Princeton.Edu	cvelasco@princeton.edu
URL:	http://imai.princeton.edu	

2 Logistics

- Lectures: Mondays and Wednesdays, 9:30am – 11:00am, 127 Corwin Hall
- Precepts: Thursdays, 4:30pm – 6:00pm, TBD
- Kosuke's office hours: Wednesdays, 11:00am – noon; or stop by anytime
- Carlos' office hours: Tuesdays, 4:00pm – 6:00pm

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 to help each other understand the materials. Carlos 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 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 a primary goal of social science research.

5 Prerequisites

There are two prerequisites for this course:

- Probability and statistics covered in POL 571: DeGroot and Schervish (2002).
- Statistical programming covered in the statistical programming camp held at the end of January. The camp materials are posted at Blackboard.

In addition, the following is strongly recommended:

- Mathematics covered in POL 502 : Basic real analysis, calculus, and linear algebra.

6 Course Requirements

The final grades are based on the following items:

- **Participation** (10%): The level of engagement in lectures, precepts, and Blackboard discussions.
- **Problem sets** (30%): Several problem sets will be given throughout the semester. Each problem set will equally contribute to the final grade and contain both analytical and data analysis questions. The following instructions will apply to all problem sets:
  - *Submission policy.* Neither late submission nor electronic submission will be accepted unless you obtain a prior approval from the instructor.
  - *Collaboration policy.* Collaboration is allowed but your study group should be no greater than three people for any given problem set. If you get help from others, you should acknowledge that in your answer. Please note that you should never *copy* someone else's answers or computer code. In particular, sharing a paper or electronic copy of your code and answers directly with other students is strictly prohibited.
- **Inclass exams** (30%): Two closed-book inclass exams given immediately after the spring break and during the first week of May, covering the first and second half of the course materials, respectively. Each exam equally contributes to the final grade.
- **Take-home final exam** (30%): The take-home open-book final exam will be given on May 9 and be due on May 16. The exam consists of data analysis questions.

7 How to Get Most out of this Course

To get most out of this course, students must keep up with the new materials that will be introduced each week by doing the following.

- Take a look at the assigned readings before coming to the lectures.
- Go over each lecture slide carefully and try to understand every detail. Forming a small study group for this is a good idea. Help others understand the materials.
- Attend precepts and use office hours and discussion board to clarify any questions you may have about the course materials.

- Start working on the problem sets as soon as you receive them. Try to solve questions on your own first before meeting with others to discuss them.
- Carefully, go over the graded problem sets and exams as well as their solutions so that you understand your mistakes.

Do not fall behind and leave any questions unanswered! 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, POL572 is a "statistics bootcamp" where you are introduced to the fundamentals of applied statistics and data analysis. In POL573, you will begin to learn how to conduct original research using statistics and data analysis. Thus, you should take POL573 in order to get most out of POL572. This means that you should carefully decide whether POL572 (and POL573) offers the training you want to receive.

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 (and 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 strongly 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.

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

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

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

David A. Freedman. *Statistical Models: Theory and Practice.* Cambridge University Press, Cambridge, 2nd edition, 2009.

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

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

Charles F. Manski. *Identification for Prediction and Decision.* Harvard University Press, Cambridge, MA, 2007.

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

Paul R. Rosenbaum. *Design of Observational Studies.* Springer, New York, 2009.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data.* The MIT Press, Cambridge, MA, 2nd edition, 2010.

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 two to three weeks on each topic but the plan is subject to change, depending on how students are keeping up with the course materials.

Basic Principles of Statistical Inference

1. Descriptive, Predictive, and Causal inference

   Lecture notes "Statistical Framework of Causal Inference"

   Freedman (2009) Chapter 1.

2. Identification, Estimation, Confidence Interval, and Hypothesis Testing

   Lecture notes: "Classical Approaches to Statistical Analysis of Randomized Experiments"

   Manski (2007) Introduction and Chapter 7.

   DeGroot and Schervish (2002) Sections 7.1–7.5, 8.1, 8.5–8.7, 9.3–9.5.

3. Problem Sets: Sample surveys, Randomized experiments

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. Matching and Regression

   Lecture notes: "Selection Bias in Observational Studies"

   Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, Summer 2007.

4. Fixed Effects, First Differences, and Difference-in-Differences

   Lecture notes: "Causal Inference with Repeated Measures in Observational Studies"

   Angrist and Pischke, Chapter 5.

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

Structural Equation Modeling

1. Instrumental Variables

   Lecture notes: "Randomized Experiments with Noncompliance"

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

   Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, 91(434):444–455, 1996.

   Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Jounal of Econometrics*, 142(2):615–635, February 2008.

2. Direct and Indirect Effects

Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789, November 2011.

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

Maximum Likelihood and Regression Models

1. Likelihood Theory

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

   David A. Freedman. On the so-called "Huber sandwitch estimator" and "robust standard errors". *American Statistician*, 60(4):299–302, 2006.

2. Bootstrap and Monte Carlo Approximation

   Freedman (2009) Chapter 8.

   Gary King, Michael Tomz, and Jason Wittenberg. Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*, 44(2): 341–355, 2000.

3. Discrete Choice Models

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

4. Sample Selection Models and Missing Data

   James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1): 153–161, January 1979.

   Gary King, James Honaker, Anne Joseph, and Kenneth Scheve. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1):49–69, March 2001.

5. Problem Sets: Retrospective sampling design, Multiple imputation