Monte Carlo Simulation and Resampling Methods for Social Scientists

Thomas M. Carsey (carsey@unc.edu)
Jeffrey J. Harden (jjharden@unc.edu)
Room 14, Manning Hall
UNC-Chapel Hill
June 27 – July 1, 2011

(Taught as part of the Odum Institute for Research in Social Science/ICPSR Summer Program)

Course Description

Social scientists increasingly use statistical simulation techniques to help them understand the social processes they care about and the statistical methods used to study them. This class will examine two types of computer simulation techniques that are quickly becoming essential tools for empirical social scientists: Monte Carlo simulation and resampling methods. We will focus on how these techniques can be used to evaluate statistical models and the resulting implications for substantive theory. A Monte Carlo simulation draws multiple samples of data based on an assumed Data Generating Process (DGP). Then the researcher explores patterns that emerge across those samples. Most importantly, the researcher controls all aspects of the DGP, which allows for precise comparison of competing theoretical models and/or statistical estimators. Resampling techniques are similar in that they also draw multiple simulated samples. However, rather than making draws from a theoretical DGP defined by the researcher, resampling techniques draw multiple simulated samples from the researchers actual sample of data. Common examples we will cover include bootstrapping, permutation and randomization testing, posterior sampling, and cross-validation. As with simulation, these methods allow researchers to evaluate both substantive theory and statistical methods.

This course will be applied in nature, with students working numerous examples. It is geared for empirical social scientists as well as students interested in studying research methods. The course will emphasize both numerical and graphical methods of evaluating simulation results. These methods provide students with a deeper understanding of key concepts such as DGPs, populations, sampling, and statistical estimators. This greatly facilitates making connections between substantive theory and empirical evidence.

Prerequisites

The course will be conducted using the programming environment `R` . Familiarity with `R` will be very helpful, but is not required. Lab computers will be available, but students wishing to work on their own laptop should download and install `R` **prior to the first course meeting**. The necessary information to do so is can be found online at: `http://cran.r-project.org/` which is the website for the Comprehensive R Archive Network (CRAN). Whether you install R on your own computer or not, **all students** are expected to review the document An Introduction to R (html or pdf version) **prior to the first course meeting**. This is also available on the CRAN website under the "Manuals link located on the bottom left of the main page.

Familiarity with basic multiple regression will be assumed. Familiarity with Maximum-likelihood estimation and models for categorical/limited dependent variables we be helpful, but not required. This course will not explore Bayesian methods or cover Markov Chain Monte Carlo (MCMC) methods.

Course Reading Materials

There is no single book that covers this material effectively. As a result, primary material will be provided in class in the form of slides and lecture notes. Much of that material will rest on the readings denoted below as "Supplemental Reading. We do NOT recommend purchasing these materials before the class. Rather, we suggest students focus on the material assigned in the class and then explore these (and other) supplemental materials later. Students looking for something to read in advance of the class should at the first two short books authored/co-authored by Mooney. Both are from the Quantitative Applications in the Social Sciences series by SAGE press (the little green SAGE books). Both provide quite accessible introductions, though the *Monte Carlo Simulation* book is written based on GAUSS rather than R .

In addition, several articles and/or book chapters will be assigned as part of the course. Some will focus on explaining a particular technique while others will simply illustrate applications of a given technique to a particular social science question. Since we are political scientists, many of our examples will come from that literature. However, we will seek to provide examples from other disciplines as well.

Supplemental Reading

Monte Carlo Simulation (1997) by Mooney.

Bootstrapping: A Nonparametric Approach to Statistical Inference (1993) by Mooney and Duval

Statistical Computing with R (2008) by Maria L. Rizzo.

Introduction to Scientific Programming and Simulation Using R (2009) by Jones, Maillardet and Robinson.

A First Course in Statistical Programming with R (2007) by Braun and Murdoch.

Introducing Monte Carlo Methods with R (2010) by Robert and Casella.

Resampling Methods: A Practical Guide to Data Analysis (2006) by Good

Introduction to Statistics Through Resampling Methods and R/S-Plus (2005) by Good.

Data Analysis by Resampling: Concepts and Applications (2000) by Lunneborg

Daily Course Schedule

What follows is a detailed schedule for the course. We intend to stick to the schedule, but we may discover that some topics take more time than anticipated and others less. We will present and discuss core material in the mornings, then spend the afternoons working with these tools while performing actual simulations and replication analyses. In other words, the afternoons will focus on “getting our hands dirty. Any additional reading beyond class handouts assigned for the day is best read **before** the start of class that day.

Day 1: Getting Started

Morning

- Course overview
- Why use simulations?
- Randomness and probability
- Begin Introduction to `R`

Afternoon

- Basic rogramming in `R`
- Probability in `R`
- Random number generation in `R`
- Running a Monte Carlo simulation

Day 2: Monte Carlo, Part 1

Morning

- Conceptual understanding of simulation
- Simulation as a controlled experiment
- Characteristics of simulations worth capturing

Afternoon

- Monte Carlo simulation and the linear model
- Simulating data with known problems
- Accessing the performance of estimators

Day 3: Monte Carlo, Part 2

Morning

- Simulating data beyond the simple linear model
- The broader use of simulation in social science

Afternoon

- Simulations on complex data types
- Presentation of simulation results

Day 4: Resampling Methods 1

Morning

- Conceptual understanding of resampling
- Common resampling methods

Afternoon

- Do-it-yourself resampling in `R`
- Resampling packages in `R`

Day 5: Resampling Methods 2

Morning

- Posterior simulation
- Cross-validation
- Brief introduction to MCMC (time permitting)

Afternoon

- Posterior simulation in `R`
- Cross-validation in `R`