Syllabus for Ec/SS 124:

Identification Problems in the Social Sciences

Winter 2014

Room: Baxter 127
Days and Times: Tuesdays and Thursdays, 10:30am - 11:55am

Instructor: Bob Sherman
Office: 119 Baxter
Phone: 4337
Office Hours: Fridays, 1:00pm - 2:30pm and by (dis)appointment

Secretary: Sabrina Boschetti
Office: 123 Baxter
Phone: 4228

Text: Manski (1995) Identification Problems in the Social Sciences, Harvard University Press.

Course Description: Statistical inference in the social sciences is a difficult undertaking whereby we combine data and assumptions to draw conclusions about the world we live in. We then make decisions, for better or for worse, based on these conclusions. A simultaneously intoxicating and sobering thought!

So, the inferences we make about the world are based on data and assumptions. Now, once observed, data is fixed. However, the assumptions we then choose to make about the data generating process can be as varied as there are theories of human behavior, and can lead to very different conclusions. Sometimes, two conflicting theories can lead to polar opposite but logically consistent interpretations of the same data set!

Even within a given theory of human behavior, the strength of the assumptions we make about the data generating process can range from very strong to very weak. In principle, the stronger our assumptions, the stronger the conclusions we can draw about things that interest us. We all like strong conclusions, but making strong assumptions to reach them comes at a price: the stronger our assumptions, the less credible our inferences. This seems particularly true in the social sciences. Often, social scientists make very strong assumptions about the data generating process. This can lead to strong and sometimes striking, but sometimes less than credible (perhaps incredible?) inferences.

In a sense, this class will err at the opposite extreme, in that we will consider a so-called "worst-case" approach to inference, popularized in recent years by Charles Manski. This approach makes the weakest, or the most conservative, assumptions, and then determines what can still be learned about things that matter. Sometimes this will imply that we cannot identify certain interesting characteristics of the underlying data distribution. At the same time, we may be able to identify informative bounds on such characteristics. In fact, you might say that this course is an introduction to bounds identification (and estimation).

Manski has long argued that a balanced approach to inference is best, whereby strong and weak (as well as "streak") approaches to assumption-making are considered and compared, to give the social scientist a clearer sense of how inferences can change as a function of assumption strength. This class will introduce some of these ideas in the context of applications of interest to social scientists.

Grading: Grades will be determined by performance on 3 homework sets, and either a project or a final exam. You choose between the project or final. Each of the 4 inputs is worth 25% of your final grade. You must turn in all four parts (3 homeworks and project or final) in order to get a passing grade. The project will involve applying the bounds techniques learned in the class to a data set of your own choosing. Naturally, it is preferable to use a data set related to a subject or question that you really care about, but you may also use a data set related to a subject or question that you couldn't possibly care less about, provided it lends itself to bounds estimation. The last day of class is Tuesday, March 11. The project will be due on this momentous day. The final will be posted to Moodle by 9am on Wednesday, March 12, and will be due under my door (Baxter 119) by 5pm on Friday, March 14.

Final letter grades will be determined as follows:

A: 90% or higher

B: 80-89%
C: 70-79%
D: 60-69%
F: Below 60%

Collaboration Policy: You are encouraged to freely collaborate on the homework sets. That is, you may discuss and even work out solutions to the homework sets together. However, you must write up your own solutions.

Moodle: I will post notes, homeworks, journal articles, etc, at Moodle, which can be accessed at

https://courses.caltech.edu/login/index.php

with your IMSS username and password, using my last name "sherman" as the enrollment key.

some book references:

1. *Manski (2013) Public Policy in an Uncertain World, Harvard
2. *Manski (2007) Identification for Prediction and Decision, Harvard
3. *Manski (2003) Partial Identification of Probability Distributions, Springer
4. Amemiya (1985) Advanced Econometrics, Harvard
5. Rao (1983) Nonparametric Functional Estimation, Academic Press
6. Hardle (1990) Applied Nonparametric Regression, Cambridge
7. King (1997) A Solution to the Ecological Inference Problem, Princeton

* means on reserve at Millikan Library (inquire at the circulation desk)

some relevant paper references:

1. Cross and Manski (2002), Regressions, Short and Long, Econometrica, 70, 357-368.
2. Cross and Manski (1999), Regressions, Short and Long, Northwestern University, manuscript.
3. Rosenbaum and Rubin (1983), The central role of the propensity score in observational studies of causal effects, Biometrika, 70, 41-55.
4. Horowitz and Manski (1995), Identification and robustness with contaminated and corrupted data, Econometrica, 63, 281-302.
5. Fan, Sherman, and Shum (2013), Partial Identification in an Ecological Inference Model, manuscript.
6. Dominitz and Sherman (2006), Identification and estimation of bounds on school performance measures: a nonparametric analysis of a mixture model with verification, Journal of Applied Econometrics, 21, 1295-1326.
7. Dominitz and Sherman (2004), Sharp bounds under contaminated or corrupted sampling with verification with an application to environmental pollution data, Journal of Agricultural, Biological and Environmental Statistics, 9, 319-338.

tentative schedule for the term:

Week 1: Introduction: Some Motivating Examples
Chapter 1: Extrapolation
Week 2: Chapter 2: The Selection Problem
Week 3: Heckman: Selection on Unobservables
Rosenbaum and Rubin: Selection on Observables (Propensity Score Methods)
Week 4: Chapter 3: The Mixing Problem in Program Evaluation
Week 5: Chapter 4: Response-Based Sampling
Week 6: Chapter 5: Predicting Individual Behavior
Week 7: Chapter 6: Simultaneity
Week 8: Chapter 7: The Reflection Problem
Week 9: Mixture Models with Verification
Love Canal
Test Scores
Ecological Inference
Week 10: Last day of class: Wrap-up (ice cream and games)

Milestones (millstones?):

January 24: add day
February 5-11: midterms
February 26: drop day
March 11: last day, projects due
March 12: final posted
March 14: finals due