EMPIRICAL STRATEGIES IN APPLIED ECONOMETRICS

Institute for the Study of Labor
April 2002

Summer School in Labor Economics
J. Angrist

This short-course covers topics in econometrics and discusses empirical modeling strategies that I have found especially useful for applied research. The main theoretical ideas are illustrated with examples.

OUTLINE AND READINGS

I. AGNOSTIC REGRESSION


II. CAUSAL REGRESSION AND REGRESSION VS. MATCHING


III. ESTIMATING THE EFFECT OF TRAINING PROGRAMS


IV. INSTRUMENTAL VARIABLES

A. Models with constant effects; Wald, grouping, and two-sample IV


B. IV with heterogeneous potential outcomes


V. MISCELLANEOUS TOPICS

A. Limited dependent variables and quantile treatment effects


B. Clustering and the Moulton problem


C. The propensity score paradox


EMPIRICAL STRATEGIES IN APPLIED ECONOMETRICS

Institute for the Study of Labor

Summer School in Labor Economics

April 2002

J. Angrist

Agenda

I. Agnostic regression – a place to start

  Regression as Best Linear Predictor (BLP) for the Conditional Expectation Function (CEF)
  Review of large sample theory for OLS estimates

II. Causal regression (our main occupation); regression vs. matching

  Linking a regression model with a causal model
  The experimentalist view of non-experimental research
  Matching to estimate the effect of treatment on the treated
  Theoretical comparison of regression and matching
  The Angrist (1998) study of the effects of voluntary military service

III. Estimating the effect of training programs – the mother of econometric evaluation problems

  Why training programs are hard to evaluate
  The Ashenfelter and Card (1985) training evaluation
  The credibility of non-experimental training evaluations; Lalonde (1986)
  Use of the propensity score in evaluation research
  The Dehejia and Wahba (1999) propensity-score study

IV. Instrumental variables (nature’s stream of experiments)

  A. Constant-effects models

     IV and omitted variables bias: long regression without the controls
     The Wald estimator and grouped data
     Two-sample IV and related methods
     The Angrist (1990) study of the effects of Vietnam-era military service

  B. IV with heterogeneous potential outcomes

     Local average treatment effects; internal vs. external validity
     The compliers concept; identification of effects on the treated
     Models with variable treatment intensity; examples
     The Angrist and Krueger (1991) schooling study

V. Miscellaneous topics

  A. Limited dependent variables and quantile treatment effects
  B. Clustering and the Moulton problem
  C. The propensity score paradox
1. Discuss the relationship between regression and matching, as described below:

   a. Suppose all covariates are discrete and you are trying to estimate a treatment effect. Prove that if the regression model for covariates is saturated, then matching and regression estimates will estimate the same parameter (i.e., have the same $plim$) in either of the following two cases: (i) treatment effects are independent of covariates; (ii) treatment assignment is independent of covariates.

   b. Propose a weighted matching estimator that estimates the same thing as regression.

   c. Why might you prefer regression estimates over matching estimates, even if you are primarily interested in the effect of treatment on the treated?

   d. (extra credit) Calculate matching and regression estimates in the empirical application of your choice. Discuss the difference between the two estimates with the aid of a figure like the one used in Angrist (1998) for this purpose.

2. Discuss the link between causal effects and structural parameters in a Bivariate Probit model of the relationship between divorce and female labor force participation. The purpose of the model is to determine whether female employment strengthens a marriage or encourages divorce more by making it easier for women to live independently.

Organize your discussion as outlined below:

   a. Explain in words why the causal effect of employment on divorce is difficult to determine. Is the problem here primarily one of identification or estimation? Could you design an experiment to answer the question of interest?

   b. Write the potential outcomes and potential treatment assignments in your causal model in terms of latent indices with unobserved random errors in a structural model.

   c. What should the population be for this study? What does it mean for employment to be “endogenous” in the structural model? How about in the causal model?

   d. Show how to use the Probit structural parameters and distributional assumptions to calculate LATE, the population average treatment effect (ATE), and the effect of treatment on the treated (ETT). Which of these parameters are identified without distributional assumptions?

   e. Discuss the relationship between the three average causal effects, LATE, ATE, and ETT. Can you say which is likely to be largest and which is likely to be smallest?

   f. (extra credit) Compare OLS with Probit and IV with Bivariate Probit in the application of your choice (as in Angrist, 2001).