Demonstration Of Instrumental Variables And Control Function Methods

Nick Mader
Chapin Hall, University of Chicago

Evan Misshula
Criminal Justice, CUNY Graduate Center

Abstract

Selection bias affects many of the most promising solutions to social problems. There are ways to correct for these risks to validity. This paper discusses Instrumental Variables and Control Functions. Specifically, This code is a simple demonstration of the use and implementation of Instrumental Variables (IV) and Control Function estimation procedures.

1 Quotes

To parents who despair because their children are unable to master the first problems in arithmetic I can dedicate my examples. For, in arithmetic, until the seventh grade I was last or nearly last. Jacques Salomon Hadamard (1865-1963)

To many, mathematics is a collection of theorems. For me, mathematics is a collection of examples; a theorem is a statement about a collection of examples and the purpose of proving theorems is to classify and explain the examples. John B. Conway

2 Introduction

To illustrate the benefits of IV and control functions, a well chosen example is worth far more than a theoretical proof. So with apologies to David Hilbert, we will choose a model designed to evoke a common problem.

3 Model

Let us start with a nested theoretical model:

$$\begin{aligned} y &= [1, x_1, x_2] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon_w \\ x_1 &= \mathbb{1}_{[-1+\beta_4 \cdot x_4 + \beta_5 \cdot x_5 + \epsilon_j]} \end{aligned}$$

Here the tuple in the first of these equations $[1, x_1, x_2]$ is our constant variable, two value categorical variable and continuous variable. These variables comprise our input data which have some unknown linear relationship to an output variable y. The error is denoted by e_w where w is the index of our principal data records of interest. The second equation, is an indicator function which predicts the value of the first of our inputs. Here ϵ_j represents our error in predicting the value of the categorical variable in the first equation. For computational purposes we can instantiate our model, using names for our hypothetical variables for familiarity rather than generalizability, is:

```
      wage = 5 + 1.0*JobTraining + 0.3*Act_Rel + e_w
JobTraining = 1[-1 + 0.2*Act_Rel - 0.4*Dist_Mi + e_j > 0]
```

In this case we are evaluating a jobs program where we believe that is having an effect but doubters point out that our results have selection bias since motivated job seekers are more likely to take the course in the first place. In many cases it is critically important for promising programs to use data to prove their positive effects are not wholly attributable to selection bias. In some cases randomized control trials (RCTs) may randomize the bias over both the control and treatment groups which will allow for unbiased estimation under ordinary least squares (OLS). However, in this case taking the course is a function of distance to the job training center so `e_w` and `e_j` are correlated since we can't control for motivation, and since motivation will determine both wages and the value of JobTraining. Only by randomizing the number of "motivated" job seekers in the training program can we estimate the actual effect of the training course. Note that, for interpretive simplicity, "Act_{Dev}", is interpreted as the relative difference between one's Act score and the average Act.

Notes on control function approach:

$$
\begin{aligned}
E[w|X,Z,D] &= E[XB + aD + e_w|X,Z,D] \\
&= XB + aD + E[e_w|X,Z,D] \\
&= XB + aD + D * (E[e_w|X,Z,D=1]) \\
&\quad + (1-D) * (E[e_w|X,Z,D=0]) \\
&= XB + aD + D * (E[e_w|e_j > -ZG]) \\
&\quad + (1-D) * (E[e_w|e_j < -ZG])
\end{aligned}
$$

See `http://en.wikipedia.org/wiki/Inverse_Mills_ratio` for constructions of thesee expectations terms.