Instrumental Variables

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Instrumental Variables

Introduction

- Instrumental variables allow us to get a better estimate of a causal treatment effect
- Their use involves an interesting algebraic "trick" that seems almost magical
- However, of course, we recognize that there is no magic in statistics — some serious assumptions are involved

Omitted Variable Bias Revisited

- Suppose the "correct" specification for confounding covariate x_i is

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i \tag{1}$$

- We are trying to assess the "value of education,' i.e., the causal impact of years of schooling on income.
- However, there is a confounding variable called ability!
- Smart people tend to do better at things, regardless of their education. Smart people also seek more education much of the time.
- Moreover, suppose that the regression for predicting x_i from the treatment is

$$x_i = \gamma_0 + \gamma_1 T_i + \nu_i$$

Omitted Variable Bias Revisited

- Substituting, we get

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 T_i + \beta_2(\gamma_0 + \gamma_1 T_i + \nu_i) + \epsilon_i \\
&= \beta_0 + \beta_2\gamma_0 + \beta_1 T_i + \beta_2\gamma_1 T_i + (\epsilon_i + \beta_2\nu_i) \\
&= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1) T_i + (\epsilon_i + \beta_2\nu_i)
\end{aligned}
\tag{2}
$$

- Note that this can be written as

$$
y_i = \beta_0^* + \beta_1^* T_i + \epsilon_i^*
$$

where

$$
\beta_1^* = \beta_1 + \beta_2\gamma_1
$$

Instrumental Variable Algebra

- An instrumental variable Z for X is a variable that is correlated with our schooling variable T, but is uncorrelated with the regression error ϵ and is also uncorrelated with the omitted variable X (ability).
- Suppose such a variable existed. What would be its properties, and how could it help us?

Instrumental Variable Algebra

- Let's assume without loss of generality that all variable are in deviation score form, and that Z is standardized.
- We are trying to estimate β_1, but the omitted variable X has made any estimate of that biased. However, I assert that

$$\beta_1 = \text{Cov}(Y, Z) / \text{Cov}(T, Z) \tag{3}$$

- Can you prove this? (C.P.)

Complicating Factors

- In many situations, you have several potential predictors and several omitted confounding variables.
- In order for the system to work, you need at least as many instruments as you have "troublesome" predictors.

Complicating Factors

- Instrumental variables seem almost magical. You might be tempted to think of them as a panacea.
- However, they can be suboptimal in several ways.
  1. The instrument can be correlated with the error term. Such an instrument is called "invalid."
  2. In his 2006 review of instrumental variables, Murray says that "all instruments arrive with the dark cloud of invalidity hanging overhead. This cloud never goes entirely away, but researchers should chase away as much of the cloud as they can." (p. 114)
  3. The instrument may be only weakly correlated with the troublesome predictor of interest.
- Murray, in two papers I have added to our reading list, gives an extensive and very entertaining discussion of instrumental variables methods in practice.

Two-Stage Least Squares Estimation

- This strategy works whether you have one troublesome predictor or several. We'll illustrate it here with one troublesome predictor.
- At stage 1, you predict the troublesome predictor T from the instrument Z using ordinary least squares, thereby obtaining predicted scores $\hat{T}$.
- At stage 2, you predict the criterion Y from $\hat{T}$, again using ordinary least squares regression.
- The β weight for predicting Y from $\hat{T}$ is the same as the weight for predicting Y from T with the omitted variable X in the prediction equation.
- Can you prove this? (C.P.)