The Observed Choice Problem in Estimating the Cost of Policies

November 7, 1997

Eric Rasmussen

Abstract

A policy will be used more heavily when its marginal cost is lower. In a regression setting, this can mean that the equation to be estimated is actually $y_i = \beta_i x(\beta_i)$. The analyst who treats times and places as identical will underestimate the policy's average cost. OLS is biased towards small coefficients, and instrumental variables should be used.

Indiana University, Kelley School of Business,BU 456, 1309 E 10th Street, Bloomington, Indiana, 47405-1701. Office: (812) 855-9219. Fax: 812-855-3354. Email: Erasmuse@indiana.edu. Web: Php.indiana.edu/~erasmuse. Copies of this paper can be found at Php.indiana.edu/~erasmuse/@Articles/Unpublished/mchoice.pdf.

It is common to estimate policy effects by looking at data from various locations. Suppose $Impact = \beta \cdot Policy$, or

$$y_i = \beta x_i, \tag{1}$$

and that the impact is undesirable. In this setting, $x_i = x(\beta_i)$ because policies are chosen in recognition of their marginal impacts in particular locations, and β varies across locations. This causes a predictable bias in OLS estimation which I call " the observed choice problem". This problem has not been directly discussed in the econometrics literature. The closest I have found is Garen (1984). In my own Rasmusen (1996) I develop the problem more fully and apply it to the slightly more complicated case where the policy impact is desirable.

The following three-equation model illustrates the bias.

$$y_i = \beta_i x_i + \epsilon_i \tag{2}$$

$$\beta_i = \overline{\beta} + v_i \tag{3}$$

$$x_i = \gamma_1 + \gamma_2 \beta_i + \gamma_3 z_i + u_i \tag{4}$$

Assume that: (i) $\gamma_1 + \gamma_2 \overline{\beta} + \frac{\gamma_3 \sum z_i}{N} > 0$, (ii) $\overline{\beta} > 0$, (iii) z and $\overline{\beta}$ are nonstochastic, (iv) ϵ, u and v are independent stochastic disturbances with mean zero and finite variance, (v) v has a symmetric distribution, (vi) $\gamma_2 < 0$. Assumptions (i) and (ii) are just normalizations, but (vi) represents that y is an undesirable impact of x, so x is used less when β_i is greater.

The OLS estimate of $\overline{\beta}$ is

$$\hat{\beta}_{OLS} = \frac{\sum x_i y_i}{\sum x_i^2}, \tag{5}$$

which has the expectation

$$E\left(\frac{\sum x_i(\overline{\beta} x_i + v_i x_i + \epsilon_i)}{\sum x_i^2}\right) = E\left(\overline{\beta}\frac{\sum x_i^2}{\sum x_i^2}\right) + E\left(\frac{\sum x_i^2 v_i}{\sum x_i^2}\right) + E\left(\frac{\sum x_i \epsilon_i}{\sum x_i^2}\right) . \tag{6}$$

The first and last terms of (6) equal $\overline{\beta}$ and 0, and the middle term equals 0 if $E(x_i^2 v_i) = 0$. If x_i and v_i are independent, OLS is unbiased.

This model, however, violates the OLS assumptions in two ways, each harmless by itself, but bad in combination: random parameters and stochastic regressors. The simpler system of just (2) and (3) has random parameters, and the simpler system of just (2) and (4) (so $\beta_i = \overline{\beta}$) has stochastic regressors, but in each of those two simple systems, OLS would be unbiased.

To see that the OLS estimate of $\overline{\beta}$ is biased in the full system, combine equations (3) and (4) to get

$$x_i = \gamma_1 + \gamma_2\overline{\beta} + \gamma_2 v_i + \gamma_3 z_i + u_i \ . \tag{7}$$

The critical middle term in equation (6), which for unbiasedness must equal zero, can be written using (7) as

$$\frac{\sum(\gamma_1 + \gamma_2\overline{\beta} + \gamma_2 v_i + \gamma_3 z_i + u_i)^2 v_i}{\sum x_i^2}. \tag{8}$$

The summed quantity in the numerator has the expectation

$$2\gamma_2[\gamma_1 + \gamma_2\overline{\beta} + \gamma_3 z_i]\sigma_v^2, \tag{9}$$

since $E(v^3) = 0$ by assumption (v), and u and v are independent.

Expression (9) has the same sign as $\gamma_2[\gamma_1 + \gamma_2\overline{\beta} + \gamma_3 z_i]$. Summed across the n observations, this takes the same sign as γ_2, since the term in square brackets is positive by assumption (i). Since $\gamma_2 < 0$, β is underestimated.

This is similar to the folk wisdom that estimation problems lead to coefficients being too small. Instrumental variables can be used to solve the observed-choice problem, as I show in Rasmusen (1996), if the analyst can observe z.

Figure 1 illustrates the problem. It shows two localities with their own relationships between policy x and impact y depicted as rays through the origin. Localities 1 and 2 have slopes β_1 and β_2, an average slope of $\overline{\beta} = \frac{(\beta_1 + \beta_2}{2}$. Policymakers 1 and 2 choose points on their respective rays. If they choose x ignoring local conditions, x_1 and x_2 have the same expected value, and the expected average of the two observations is on the middle ray. This corresponds to OLS being unbiased.

If, however, y is a cost of x, and a steeper slope makes a policymaker choose a lower level of x, then Locality 1, with a greater marginal cost, chooses a lower x than Locality 2: $x_1 < x_2$. If the econometrician draws a line through the origin to lie between the two observations and minimize the squared deviations, that line will have a slope of less than $\overline{\beta}$. OLS underestimates the marginal cost.

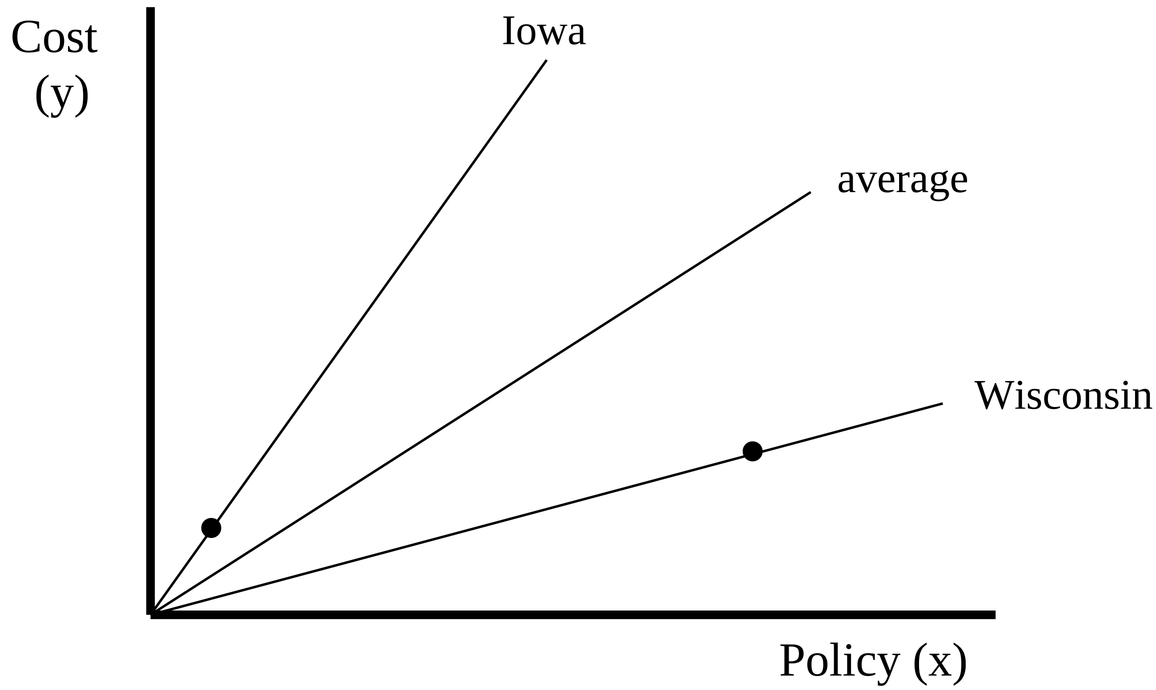

FIGURE 2: ESTIMATING THE MARGINAL COST OF A POLICY

REFERENCES

Garen, John (1984). The returns to schooling: A selectivity bias approach with a continuous choice variable. *Econometrica* 52 (September): 1199-1218.

Rasmusen, Eric (1996) "Observed Choice and Optimism in Estimating the Effects of Government Policies," forthcoming, *Public Choice*.