Instrumental variables in-class exercise

Nathaniel Higgins
nhiggins@jhu.edu

Run (or enter manually) the code to create our dset for this question (consult "lecture-10-iv.r"). You should have the following variables now in your dset: `x1, x2, x3, x4, u`, and `y`. The true model for `y` (the true *data generating process* of `y`) is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x + 2 + u. \tag{1}$$

In fact, we can tell from the code that the true data generating process of `y` is

$$y = 5 + 2x_1 - 4x_2 + u.$$

We are interested in the impact of `x1` on `y`. Suppose that `x2` is unobservable (it's something like the "ability" variable we have been using in class thus far). This forces us to look at the relationship between `y` and `x1` without controlling for `x2`, for instance by running the following regression.

$$\texttt{lm(y} \sim \texttt{x1)}$$

If `x2` were uncorrelated with `x1`, we wouldn't have a problem. Although including `x2` in a regression would help us to predict `y` better, excluding `x2` wouldn't hurt our ability to estimate the impact of `x1` on `y`. But this isn't the case. You can tell by reading the code that `x1` and `x2` are correlated. Your task is to estimate the impact of `x1` on `y` using what you have learned today.

Do the following

1. We don't have `x2`. If you were to run a regression of `y` on `x1`, would the estimated effect of `x1` on `y` be unbiased? If there is a bias, in what direction is the bias? Do you expect the estimated coefficient to be too high or too low relative to the true value? Answer this question before completing #2 below.

2. Regress `y` on `x1` and record what happens (do you get a "good" result? is the true value of the coefficient contained in the 95 % confidence interval?)

3. You have two variables other than `y` and `x1` in your dset: `x3` and `x4`. You can consider these to be your *potential* instruments in the dset. Of course, remember that `x2` is NOT available to you (and `u` is, as always, an unobservable). Which instrument will work (better)? Why? Answer this question just by looking at the code and thinking through the problem, before actually trying to carry out the regressions necessary to uncover the impact of `x1` on `y` (you will do this below). That is, answer this question as if it were an exam question — and you don't have Stata on the exam!

4. Now, try instrumenting for `x1`. Do this by predicting `x1` with those variables that are available to you (those variables which you considered above). Then use this prediction to uncover the relationship between `y` and `x1`. That is, test your answer to #3 above: run your instrumental variables regression(s). Is your answer to #3 validated? When you run your chose IV regression(s), is the true value of the coefficient in the 95% confidence interval?