Problem Set 2

Operation IV

EC 425/525: Econometrics

Due *before* midnight (11:59pm) on Wednesday, 29 May 2019

DUE Your solutions to this problem set are due *before* 11:59pm on Wednesday, 29 May 2019 on Canvas. Your problem set **must be typed** with `R` code beneath your responses. *E.g.*, `knitr` and `R Markdown`.

OBJECTIVE We're going to walk through three classic applications of instrumental variables/two-stage least squares: endogeneity, measurement error, and randomized encouragement designs (REDs).

Part 1: Selection bias

As this problem follows one from Wooldridge, we'll use the `wooldridge` package. You need to install the `wooldridge` package and then load the birthweight data using `data("bwght")`. For (limited) information on the variables, see the help file (*i.e.*, `?wooldridge::bwght`).

`1.01` We want to better understand the effect of a number of variables on birth weight (`bwght`)—namely gender (`male`), birth order (`parity`), income (`faminc`), and cigarette smoking during pregnancy (`packs`), *i.e.*,

$$\log(\text{bwght}_i) = \beta_0 + \beta_1 \text{male}_i + \beta_2 \text{parity}_i + \beta_3 \log(\text{faminc}_i) + \beta_4 \text{packs}_i + u_i$$

`1.01` Why might you expect amount of smoking (`packs`) to be correlated with u_i?

`1.02` Suppose that you have data on average cigarette prices in each woman's state of residence. Discuss whether this information is likely to satisfy the properties of a good instrumental variable for `packs`.

`1.03` Use the data in in `bwght` to estimate equation the equation above. First, use OLS. Then, use 2SLS, where `cigprice` is an instrument for `packs`. Discuss any important differences in the OLS and 2SLS estimates.

`1.04` Estimate the reduced form for `packs`. Does it raise any issues? What bearing does this conclusion have on your answer from **`1.03`**?

Part 2: Randomized encouragement designs

Another common implementation of IV/2SLS is a randomized encouragement design (RED), in which we randomly select individuals to receive "encouragement" (*e.g.*, we call them to tell them about an exciting new program) in order to try to induce an exogenous change in program participation.

Let's imagine we want estimate the effect of solar-panel installation on household electricity consumption.

`2.01` What would be the problem with comparing average electricity consumption for houses with solar panels to average electricity consumption without solar panels?

`2.02` We randomly select 200 homes that have not yet installed solar panels. Within this sample, we randomly assign 100 houses to our "encouragement" group and 100 houses to our "non-encouragement" group. For the 100 houses in the encouragement group, we call/visit the households and tell them how awesome solar panels are—and how much money they could save with solar.[†]

What do we need for our encouragement to be a valid instrument for solar panel installation? Do you think it is satisfied?

---

† Tangent: In case you haven't seen it, you should check out Google's Project Sunroof.

2.03 A year later, we conduct a survey and find that in the encouragement group, 15/100 homes now have solar panels. In the non-encouragement group, 5/100 homes now have solar panels. If we estimated the first stage, (regressing an indicator for solar panel on an intercept and an indicator for encouragement group), what would our estimates be?

2.04 Imagine that average monthly electricity consumption in the encouragement group is 900 kWh (kilowatthours), while the average in the non-encouragement is 870 kWh. Based upon these numbers, what are the reduced-form (the effect of encouragement on energy consumption) and 2SLS estimates?

2.05 What does the LATE in this setting mean—*i.e.*, what does *local* mean in this setting?

Part 3: Measurement error

Now for a good, old-fashioned simulation.

3.01 Set up a data-generating process such that

$$y_i = 3 + 7x_i + u_i$$

where $x_i \overset{\text{iid}}{\sim} N(5,5)$ and $u_i \overset{\text{iid}}{\sim} N(0,3)$.

In this simulation, we want to imagine what would happen if we could not observe x_i (or if x_i is measured with error/noise).

Thus, we want to create two additional variables: w_{1i} and w_{2i}, such that

$$\begin{aligned} w_{1i} &= x_i + \varepsilon_i \\ w_{2i} &= x_i + \nu_i \end{aligned}$$

where ε_i is i.i.d. standard Normal and ν_i is i.i.d. $N(0,7)$. For this problem, the sample size will be 50.

This setting is *classical* measurement error—the error (or noise) in measurement (*i.e.*, ε_i and ν_i) is uncorrelated with the true variable (x_i).

Note: No results for this part of the problem. Just make sure you've set up the DGP.

3.02 Imagine you cannot observe x_i and are stuck with our noisily measured versions w_{1i} and/or w_{2i}. Regress y_i on w_{1i}. What do you get? What if you regress y_i on both w_{1i} *and* w_{2i}?

3.03 Now what happens if you *instrument* w_{1i} with w_{2i}?

3.04 Confirm your results form **3.02** and **3.03** were not anomalies. In other words, run a simulation (with at least 1,000 iterations). In each iteration, record the results of

- regressing y_i on w_{1i}
- regressing y_i on w_{2i}
- instrumenting w_{1i} with w_{2i}
- instrumenting w_{2i} with w_{1i}

Report the results of your simulation. Do you see anything interesting? Does IV outperform OLS in the presence of measurement error (in terms of bias in $\hat{\beta}_1$)? What happens in your inference (look at the share of estimates in which you reject the null)?

3.05 Now let x_i **positively** correlate with ε_i and **negatively** correlate with ν_i, *i.e.*, $\text{Cov}(x_i, \varepsilon_i) = 1$ and $\text{Cov}(x_i, \nu_i) = -2$. What happens to your results from **3.04**?

Hint You can use `mvrnorm()` from `MASS` to draw correlated variables from a multivariate Normal distribution (which you can assume here). See our simulation lab for details.

Extra credit For our simple linear regression setup, show (analytically) why OLS estimates for β_1 are biased toward zero. How does IV help?