Problem Set: IV and Natural Experiments

Consider the following structural equation, which relates an outcome variable y with a treatment variable x:

$$y_i = \beta x_i + \epsilon_i$$

where i denotes observations associated with different economic agents i.

Below, I give you several variables z. For each z, give:

- An example of an outcome y and treatment x for which the variable z can be used as an instrumental variable in estimating the treatment effect β.
- For the y and x you give, give one reason why x is endogenous (i.e., why x is correlated with ϵ).
- Rigorously justify why z is a valid instrument (make sure it satisfies all the assumption that an IV must satisfy).
- If the treatment effect β_i varied across agents i, the resulting linear IV estimate yields the average treatment effect for a particular subpopulation given an additional monotonicity assumption. State the monotonicity assumption for the given (y, x, z), and also state the subpopulation which the estimate is applicable to.
- For the (y, x) that you give, suggest *another instrument* that can be used to estimate β, and justify it.

1. The Vatican II council changed the language of the Catholic services from Latin to the vernacular (everyday) language. $z_t = 0$ before the council, and $z_t = 1$ afterwards.
2. $z_i = 1$ if household i lives in a rural setting (i.e., in the countryside), and $z_i = 0$ if household i lives in a city.
3. $z_i = 1$ if household i lives above the 5th floor of a high-rise apartment building, and $z_i = 0$ if above the 5th floor.
4. $z_i = 1$ if agent i's language can be written in the Roman alphabet, and $= 0$ if not.
5. $z_i = 1$ if agent i has a last name in the range A–M, and $= 0$ otherwise.

6. $z_i = 1$ if agent i is left-handed, $= 0$ otherwise.
7. $z_i = 1$ if the street address of household i is an even number, $= 0$ otherwise.
8. $z_t = 1$ when one lane of Highway 95 is closed for repairs, $= 0$ otherwise.

■■■

The second part of this problem set is a "Monte Carlo" exercise, where you simulate some fake data, and then run some regressions in order to compare how well the results from the empirical procedure correspond to the "truth".

I would suggest using MATLAB to do this exercise.

1. Linear IV model

Generate some fake data on (y, p, z) as follows:

- Let $T = 100$ observations
- For each observation, generate $(\epsilon_t, \eta_t)' \sim N\left(0, \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}\right)$. (Use the Cholesky factorization method, described in previous week's lecture notes.)
- Generate $z_t \sim U[0, 1]$.
- Generate $\tilde{z}_t = \mathbf{1}(z_t > 0.5)$.
- Generate prices $p_t = 2z_t + \eta_t$.
- Generate demand $y_t = 10 - p_t + \epsilon_t$.
- Also generate $\tilde{y}_t \equiv y_t - \frac{1}{T}\sum_t y_t$, the deviation of demand from its mean. Also generate $\tilde{p}_t$ analogously.

(a) Use OLS to estimate the demand function, using only p as a regressor.

(b) Use 2SLS to estimate the demand function, using z as an IV. Repeat, using $\tilde{z}$ as an IV.

(c) Compute the Wald estimator $\hat{\alpha}_{1,0}$ (as denoted in the lecture notes), using the differenced variables $\tilde{y}$ and $\tilde{p}$, and using the binary instrument $\tilde{z}$.

(d) Compute the Wald estimator $\alpha(z)$ (as denoted on pp. 8-9 of the notes), again using the differenced data $\tilde{y}$ and $\tilde{p}$, but using the continuous instrument z. Compute this for $z = 0.25, 0.5, 0.75$.

Use the Epanechnikov kernel function $\mathcal{K}(u) = \frac{3}{4}(1 - u^2)\mathbf{1}(|u| \leq 1)$.

Use "finite differences" to approximate the derivative $\hat{q}'(z)$:

$$\hat{q}'(z) \approx \frac{\hat{q}(z+h) - \hat{q}(z)}{h}$$

for some small h (i.e. take $h = 0.001z$, for instance).

Compare and contrast the different estimates. Can you say anything about what the estimates *should* be?

2. Nonlinear IV Model

Next we consider an example of case 4 (from the lecture notes).

Generate some fake data on (y, p, z) as follows:

- Let $T = 100$ observations
- For each observation, generate $(\epsilon_t, \eta_t, \nu_t)' \sim N\left(0, \begin{bmatrix} 1 & 0.4 & 0.6 \\ -0.2 & 1 & -0.2 \\ 0.6 & 0.4 & 1 \end{bmatrix}\right)$
- Generate $z_t \sim U[0, 1]$.
- Generate $\tilde{z}_t = \mathbf{1}\left(z_t > 0.5\right)$.
- Generate $\alpha_t = -\max\left(|\nu_t|, 2\right)$.
- Generate prices $p_t = \min\left(2, 2\tilde{z}_t + \eta_t\right)$.
- Generate demand $y_t = 10 - \alpha_t p_t + \epsilon_t$.

(a) Compute the Wald estimator $\alpha_{1,0}$, using the un-differenced data (y, p) and the binary instrument z. What is the right interpretation of this?

(b) For each observation, calculate the two potential prices $\tilde{p}(z = 1, t)$ and $\tilde{p}(z = 0, t)$. For which observations is the estimate from the previous question relevant? What percentage of the observations is this?