Instrumental Variable

Contents

The Intuition for IV

When we study the effect of x on y, we always want to control the possible confounded. However, it often happens that some confounded are not observable. How to deal with it?

One good approach is to use instrument variable approach. From causal graph sense, if we can find a variable z s.t.z leads to x, and there is no additional path through which z affects y other than via x, then we can make use of z to get a consistent estimation of the effect of x on y.

We are already familiar with instrument variable.

$$y = \beta * x + \epsilon$$

in which $cov(x, \epsilon) \neq 0$. If we can find a z such $cov(x, z)/ne0$ and $cov(z, \epsilon) = 0$:

$$x = \gamma z + u, \quad \gamma \neq 0$$

Then we can estimate β by

$$\beta = \frac{cov(y, z)/var(z)}{cov(x, z)/var(z)}$$

Intuitively, to consistently estimate β, we first estimate the effect of z on y, i.e., $cov(y, z)/var(z)$. This is consistent, since we assume that z is exogenous. But since we are asking the effect of x on y, and since z impacts y via x, we 'scale' the effect of z on y by the effect of z on x, to get the effect of x on y.

Writing the whole in expectation form, we have

$$E(y|z) = E(\beta x + \epsilon|z) = \beta E(x|z) + E(\epsilon|z) = \beta E(x|z)$$

Again, the expression shows that we can : - regress y on z to get fitted value $\hat{y}$. - regress x on z to get fitted value $\hat{x}$. - regress $\hat{y}$ on $\hat{x}$ to get the estimation on β.

Above method requires that we run three regressions. Another more common and simple method is 2SLS. Notice that

$$y = x\beta + \epsilon = (x - E(x|z))\beta + E(x|z)\beta + \epsilon = E(x|z)\beta + v$$

where $v \equiv (x - E(x|z))\beta + \epsilon$. Of course, $cov(E(x|z), v) = 0$. Therefore we can get a consistent estimation on β.

A most common application of IV in business world is the following situation: the company assigns a campaign randomly on some users $Z = 1$. The users who are assigned the $T = 1$ may choose to join in campaign $D = 1$ or not $D = 0$. The users who are assigned the $T = 0$ may also find some ways to join the campaign. We want to study the effect of D on users' behavior y. Apparently, some unobservable factors may determines both a

users' D and y. Fortunately, we have z as IV: It is closely related to D; it is random, so it affects y only through D.

Consider

$$y = x\beta + \gamma D + \epsilon$$

in which $cov(x, \epsilon) = 0, cov(D, \epsilon) \neq 0$. We can do the 2SLS:

- regress D on x and Z and get $\hat{D}$.
- regress y on x and $\hat{D}$ to get the estimation on γ.

Application of IV in panel data

Two-period DID

Consider

$$y_{it} = \alpha + \beta_1 s_{it} + \beta_2 G_t + \beta_3 D_i + \beta_4 x_{it} + \epsilon_{it}, t = 0, 1$$

At $t = 0$, no campaign happens. At $t = 1$, some users receive campaign $Z = 1$ or not. if a user i , choose to participate, then $s_{i1} = 1, D_i = 1$. Notice that $s_{i0} = 0$ always holds. x_i are some other covariates. G_t denotes time.

do difference.

$$\Delta(y_i) \equiv y_{i1} - y_{i0} = \beta s_{i1} + \beta_1 + \Delta x_i + \epsilon_{i1} - \epsilon_{i0}$$

Since this is two-period data, $s_{i1} = 1(0)$ means that user i (does not) joins the campaign. It may be correlated with $\epsilon_{i1} - \epsilon_{i0}$, leading to endogneneity. Fortunately, we have Z, which can be our IV.

- First, regress s_{i1} on Z_i and Δx_i. Get $\hat{s_{i1}}$
- Second, regress Δy_i on $\hat{s_{i1}}$ and Δx_i

Multiple-period DID

The logic is similar with the two-period DID. Consider

$$y_{it} = \alpha + \beta_1 s_{it} + \beta_2 G_t + \beta_3 D_i + \beta_4 x_{it} + \epsilon_{it}$$

we have

$$y_{it} - \overline{y_i} = \beta_1 \left(s_{it} - \overline{s_i}\right) + \beta_2 (G_t - \overline{G}) + \beta_4 (x_{it} - \overline{x_i}) + \epsilon_{it} - \overline{\epsilon_i}$$

Also consider a variable z_{it}, s.t., $z_{it} = 1$ means that the user i receives campaign at time t.

$s_{it} = 1$ when user i joins the campaign at time t.

If z_{it} is purely a random design across individuals and time, then we can use $z_{it} - \overline{z_i}$ as an instrument for $(s_{it} - \overline{s_{it}})$ and do the estimation.