A note on the instrumental variable and local average treatment effect

Yen-Chi Chen
University of Washington[1]

September 29, 2022

The instrumental variable (IV) is a powerful approach to draw causal conclusion even the data is not obtained from an experiment. It is a popular method in both Economics and Medical research and a power of IV is that it can be applied to cases when there are unobserved confounders. In this note, we will briefly review its properties.

Let Y be the outcome of interest and $A \in \{0,1\}$ denote the treatment ($A = 1$ often refers to those who receive the treatment) and $Z \in \{0,1\}$ be a binary variable (it will be called a valid IV under certain conditions). One can think of the treatment A to be the actual treatment that an individual takes and Z is the suggested treatment from the doctor. In the randomized trials, the suggested treatment is always the same as the actual treatment. But in many observational studies, these two are different quantities.

The treatment A and the variable Z induces 4 potential outcomes $Y(a,z)$: $Y(1,1),Y(1,0),Y(0,1),Y(0,0)$. The variable Z also creates two potential outcome $A(z) : A(1),A(0)$. $A(1)$ is the treatment that the individual will take if the doctor's suggestion is $Z = 1$ (take the treatment).

Our primary interest is to understand the treatment effect of A on Y. Our data is the IID triplet

$$(A_1,Z_1,Y_1),\cdots,(A_n,Z_n,Y_n).$$

Note that $A_i = Z_iA_i(1)+(1-Z_i)A_i(0)$ and $Y_i = Y_i(1,1)A_iZ_i+Y_i(1,0)A_i(1-Z_i)+Y_i(0,1)(1-A_i)Z_i+Y_i(0,0)(1-A_i)(1-Z_i)$. Thus, we only observe one of the two potential outcomes of A and one of the four potential outcomes of Y.

When there is no Z and the data is from a randomized experiment, a common quantity of the treatment effect is the average treatment effect (ATE) $\mathbb{E}(Y(A = 1)-Y(A = 0))$. However, we cannot get a precise point estimate of the ATE when the experiment is not randomized[2]. Interestingly, if the variable Z satisfies certain conditions, we are able to identify *local average treatment effect (LATE)*:

$$\theta_{\mathsf{LATE}} \equiv \mathbb{E}(Y(A = 1)-Y(A = 0)|A(1) = 1,A(0) = 0).$$

Those individuals with $A(1) = 1$ and $A(0) = 0$ are often refer to as *complier*.

In the framework of IV, we often assume the following four key conditions (slightly stronger than necessary):

1. **IV1 (exclusion restriction):** $Y(a,z) = Y(a,z')$ for all $a,z,z' \in \{0,1\}$. Namely, the IV has no effect on the outcome of interest. This implies that $Y(a,z) = Y(A = a)$ so we simply denote $Y(1) = Y(A = 1)$ and $Y(0) = Y(A = 0)$ under IV1.
2. **IV2 (unconfounded/randomization):** $Z \perp (Y(a,z),A(z') : a,z,z' \in \{0,1\})$. This is a critical condition in IV analysis but is often reasonable; going back to the example. this requires that the doctor randomly suggest treatment/control to the individual in the study.

[1] We thank Thomas Richardson for very helpful comments.

[2] One may still identify it if the confounders are all observed.

3. **IV3 (no defier/monotonicity):** $A(1) \geq A(0)$.
4. **IV4 (IV relevance):** $\mathbb{E}(A|Z=1) \neq \mathbb{E}(A|Z=0)$.
5. **IV5 (positivity):** $1 > P(Z=1) > 0$.

If Z satisfies the above conditions with respect to A and Y, it is called a valid IV.

The condition IV3 is related to the classification[3] of individual's behaviors. Depending on the potential outcome of the treatment $A(z)$, we classify individuals into 4 possible groups:

$A(1)$	$A(0)$	
1	1	a: always-taker
1	0	c: complier
0	1	d: defier
0	0	n: never-taker

Table 1: The classification of individuals, also known as the principal stratum framework.

In the classical example that A represents the actual treatment and Z represent doctor's suggestion, compliers are those who will follow the doctor's suggestion, always-takers are those who will always take the treatment regardless of the doctor's suggestion, and the never-takers are those who will never take the treatment. The defier is the case where individuals will purposely 'disobey' the doctor's suggestion. While IV3 may look odd at the first glance, it is a reasonable assumption since it only assumes that no one will purposely disobey doctor's suggestion (we allows the never-takers to exist).

With the above classification the LATE can be written as

$$\theta_{\mathsf{LATE}} \equiv \mathbb{E}(Y(A=1) - Y(A=0)|\mathsf{complier}) \tag{1}$$

so it is also called *complier treatment effect*. While it is not the ATE, LATE is still of a lot of interest since it describes the treatment effect for those who follow the doctor's suggestion (which is the majority people).

1 Estimation of LATE

A powerful feature of IV is that under assumptions IV1-4, the LATE can be written as the following form

$$\theta_{\mathsf{LATE}} \equiv \mathbb{E}(Y(A=1) - Y(A=0)|\mathsf{complier}) = \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(A|Z=1) - \mathbb{E}(A|Z=0)} \equiv \theta^*_{\mathsf{LATE}}. \tag{2}$$

Note that the last quantity in the above equation is identifiable from the data since we can easily do a conditional mean estimation of Y and A to obtain a consistent estimator.

Now we derive equation (2).

[3] Also known as the principal stratum framework in Frangakis, C.E. and Rubin, D.B., 2002. Principal stratification in causal inference. Biometrics, 58(1), pp.21-29.

A useful insight. Here is the key insight of the derivation: by properties of conditional independence, IV2 will imply

$$Z \perp (Y(1,1),Y(1,0),Y(0,1),Y(0,0))|A(1),A(0). \tag{3}$$

Note that the quantities being conditioned is the potential outcomes $A(1),A(0)$, which will be described by the classification (always-taker, complier, ...). Thus, (3) further implies

$$Z \perp (Y(1,1),Y(1,0),Y(0,1),Y(0,0))|\mathsf{C}, \tag{4}$$

where C = always-taker/complier/never-taker/defier. Note that under IV1, there is only two potential outcomes of Y so we further have

$$Z \perp (Y(1),Y(0))|\mathsf{C}. \tag{5}$$

Analysis on the numerator of (2). First, we focus on the term $\mathbb{E}(Y|Z=1)$. Let a,c,n denote the always-taker, complier, and never-taker. We can decompose it as

$$\begin{aligned}\mathbb{E}(Y|Z=1) &= \mathbb{E}(Y|Z=1,a)P(a|Z=1)+\mathbb{E}(Y|Z=1,c)P(c|Z=1)+\mathbb{E}(Y|Z=1,n)P(n|Z=1)\\ &= \mathbb{E}(Y|Z=1,a)\pi_a+\mathbb{E}(Y|Z=1,c)\pi_c+\mathbb{E}(Y|Z=1,n)\pi_n,\end{aligned}$$

where $\pi_a = P(a) = P(A(1)=A(0)=1)$ denotes the proportion of always-taker (similarly for π_c,π_n). Note that by IV2, $P(a) = P(a|Z=1)$.

Conditioned on $Z=1$, always-takers will always have $A=1$ so

$$\mathbb{E}(Y|Z=1,a) = \mathbb{E}(Y(1)|Z=1,a) \overset{(5)}{=} \mathbb{E}(Y(1)|a),$$

where the last equality follows from equation (5). Conditioned on $Z=1$, conplier will have $A=1$ so

$$\mathbb{E}(Y|Z=1,c) = \mathbb{E}(Y(1)|Z=1,c) \overset{(5)}{=} \mathbb{E}(Y(1)|c).$$

Conditioned on $Z=1$, never-takers will always have $A=0$ so

$$\mathbb{E}(Y|Z=1,n) = \mathbb{E}(Y(0)|Z=1,n) \overset{(5)}{=} \mathbb{E}(Y(0)|n).$$

Thus, putting all these together, we conclude that

$$\mathbb{E}(Y|Z=1) = \mathbb{E}(Y(1)|a)\pi_a+\mathbb{E}(Y(1)|c)\pi_c+\mathbb{E}(Y(0)|n)\pi_n. \tag{6}$$

A similar derivation also applies to $Z=0$ but note that the complier will take $A=0$ under this scenario so we have

$$\mathbb{E}(Y|Z=0,a) = \mathbb{E}(Y(1)|a), \quad \mathbb{E}(Y|Z=0,c) = \mathbb{E}(Y(0)|c), \quad \mathbb{E}(Y|Z=0,n) = \mathbb{E}(Y(0)|n),$$

and

$$\mathbb{E}(Y|Z=0) = \mathbb{E}(Y(1)|a)\pi_a+\mathbb{E}(Y(0)|c)\pi_c+\mathbb{E}(Y(0)|n)\pi_n. \tag{7}$$

Thus, the difference in equation (2) is

$$\mathbb{E}(Y|Z=1)-\mathbb{E}(Y|Z=0) = (\mathbb{E}(Y(1)|c)-\mathbb{E}(Y(0)|c))\pi_c.$$

Thus, what remains to be shown is that the denominator $\mathbb{E}(A|Z=1)-\mathbb{E}(A|Z=0)=\pi_c$.

Analysis on the denominator of (2)**.** We apply a similar derivation as the case of numerator. First, we consider $\mathbb{E}(A|Z=1)$. It can be decomposed into

$$\mathbb{E}(A|Z=1)=\mathbb{E}(A|Z=1,a)\pi_a+\mathbb{E}(A|Z=1,c)\pi_c+\mathbb{E}(A|Z=1,n)\pi_n.$$

Under $Z=1$, always-takers will always have $A=1$, compliers will have $A=1$, and never-takers have $A=0$, leading to $\mathbb{E}(A|Z=1,a)=\mathbb{E}(A|Z=1,c)=1$ and $\mathbb{E}(A|Z=1,n)=0$. So we conclude

$$\mathbb{E}(A|Z=1)=\pi_a+\pi_c.$$

By a similar argument for the case of $\mathbb{E}(A|Z=0)$, we have

$$\mathbb{E}(A|Z=0)=\pi_a.$$

Thus, indeed we have

$$\mathbb{E}(A|Z=1)-\mathbb{E}(A|Z=0)=\pi_c.$$

As a result, we conclude that

$$\underbrace{\frac{\mathbb{E}(Y|Z=1)-\mathbb{E}(Y|Z=0)}{\mathbb{E}(A|Z=1)-\mathbb{E}(A|Z=0)}}_{=\theta^*_{\mathsf{LATE}}}=\mathbb{E}(Y(1)|c)-\mathbb{E}(Y(0)|c)\equiv\theta_{\mathsf{LATE}}.$$

Remark. While we do not directly use IV4 and IV5 in the derivation, one can see that IV4 is to ensure that the denominator of θ^*_{LATE} is non-zero. Also, IV5 is to make sure that we indeed have observations that $Z=1$ and $Z=0$.

2 Testing the IV conditions

The no-defier condition (IV3) is in fact testable. Suppose we do not assume IV3, we will obtain

$$\mathbb{E}(A|Z=1)=\pi_a+\pi_c,\qquad \mathbb{E}(A|Z=0)=\pi_a+\pi_d,$$

where π_d is the proportion of defier. In the above derivation of LATE, we use IV3 to enforce $\pi_d=0$, which leads to the follow inequality constraint:

$$\mathbb{E}(A|Z=0)=\pi_a\leq\pi_a+\pi_c=\mathbb{E}(A|Z=1). \tag{8}$$

The two conditional proportions, $\mathbb{E}(A|Z=0)$ and $\mathbb{E}(A|Z=1)$, can be estimated nonparametrically (by a simple sample proportion). And equation (8) may not hold in practice.

When equation (8) does not hold, one has to be very careful about using the IV approach because the IV assumptions are conflicting with the data.

3 Efficient influence function of LATE

The quantity

$$\theta^*_{\mathsf{LATE}} = \frac{\mathbb{E}(Y|Z=1)-\mathbb{E}(Y|Z=0)}{\mathbb{E}(A|Z=1)-\mathbb{E}(A|Z=0)}$$

is an identified quantity that can be expressed as a statistical functional of the generating model. Namely, let F_0 denotes the distribution (CDF) that generates our data; we can then view θ^*_{LATE} as a statistical functional $\theta^*_{\mathsf{LATE}}(F_0)$. So we can derive its efficient influence function (EIF).

The EIF of θ^*_{LATE} is

$$\begin{aligned}\mathsf{EIF}(\theta^*_{\mathsf{LATE}}) = \frac{1}{\alpha_1-\alpha_0}\Bigg\{&\frac{Z}{\mathbb{E}(Z)}(Y-\mathbb{E}(Y|Z=1)) + \frac{1-Z}{1-\mathbb{E}(Z)}(Y-\mathbb{E}(Y|Z=0)) \\ &+ \frac{\theta^*_{\mathsf{LATE}}}{\alpha_1-\alpha_0}\left[\frac{Z}{\mathbb{E}(Z)}(A-\mathbb{E}(A|Z=1)) + \frac{1-Z}{1-\mathbb{E}(Z)}(A-\mathbb{E}(A|Z=0))\right]\Bigg\},\end{aligned} \tag{9}$$

where $\alpha_z = \mathbb{E}(A|Z=z)$.

Here we will derive the EIF of equation (9). We first review some useful tricks in deriving EIF.

Useful tricks for deriving the EIF. Let ϕ_1 and ϕ_2 be two statistical functionals. We have the following properties:

$$\begin{aligned}\mathsf{EIF}(\phi_1+\phi_2) &= \mathsf{EIF}(\phi_1)+\mathsf{EIF}(\phi_2) \\ \mathsf{EIF}(\phi_1\times\phi_2) &= \mathsf{EIF}(\phi_1)\phi_2+\mathsf{EIF}(\phi_2)\phi_1 \\ \mathsf{EIF}\left(\frac{\phi_1}{\phi_2}\right) &= \frac{\mathsf{EIF}(\phi_1)}{\phi_2} - \frac{\phi_1}{\phi_2}\frac{\mathsf{EIF}(\phi_2)}{\phi_2}.\end{aligned} \tag{10}$$

You may verify these equalities. The high-level idea of why this equalities hold is that the EIF is derived based on concepts of 'differentiation'. So these three equalities are essentially the laws of derivatives. See Section 3.4 of https://arxiv.org/abs/2203.06469v1.

Analysis of $\mu_1(F_0) = \mathbb{E}(Y|Z=1)$. With the above trick, we only need to derive the EIF of $\mathbb{E}(Y|Z=1)$. Let (p_0, F_0) denotes the original PDF/CDF that generates our observed data and $(p_\varepsilon, F_\varepsilon)$ be the perturbed PDF/CDF such that $p_\varepsilon(a,z,y) = p_0(a,z,y)(1+\varepsilon g(a,z,y))$ with $\int p_0(a,z,y)g(a,z,y)dadzdy = 0$.

Because $\mathbb{E}(Y|Z=1) = \int yp(y|z=1)dy$, its perturbed value will be

$$\mu_1(F_\varepsilon) = \int yp_\varepsilon(y|z=1)dy = \frac{\int yp_\varepsilon(y,z=1)dy}{p_\varepsilon(z=1)}. \tag{11}$$

For the numerator,

$$\begin{aligned}\int yp_\varepsilon(y,z=1) &= \int yI(z=1)p_\varepsilon(y,z,a)dzdady \\ &= \int yI(z=1)p_0(y,z,a)(1+\varepsilon g(y,z,a))dzdady \\ &= \int yp_0(y,z=1)dy + \varepsilon\cdot\int yI(z=1)p_0(y,z,a)g(y,z,a)dydzda.\end{aligned}$$

For the denominator,

$$\begin{aligned} p_\varepsilon(z=1) &= \int I(z=1)p_\varepsilon(a,z,y)dadzdy \\ &= p_0(z=1)+\varepsilon\int I(z=1)p_0(a,z,y)g(a,z,y)dadzdy. \end{aligned}$$

Thus, (11) becomes

$$\begin{aligned} \mu_1(F_\varepsilon) &= \frac{\int yp_\varepsilon(y,z=1)dy}{p_\varepsilon(z=1)} \\ &= \mu_1(F_0)+\varepsilon\cdot\int\frac{yI(z=1)}{p_0(z=1)}p_0(y,z,a)g(y,z,a)dydzda \\ &\quad -\varepsilon\frac{\mu_1(F_0)}{p_0(z=1)}\int I(z=1)p_0(y,z,a)g(y,z,a)dydzda+O(\varepsilon^2). \end{aligned}$$

Thus, the EIF of μ_1 is

$$\mathsf{EIF}(\mu_1)=\frac{I(Z=1)}{P(Z=1)}\left(Y-\mathbb{E}(Y|Z=1)\right)=\frac{Z}{\mathbb{E}(Z)}\left(Y-\mathbb{E}(Y|Z=1)\right). \tag{12}$$

Other terms. Let $\mu_0(F_0)=\mathbb{E}(Y|Z=0)$ and $\alpha_z=\mathbb{E}(Y|Z=z)$. By a similar derivations, the EIF of other terms will be:

$$\begin{aligned} \mathsf{EIF}(\mu_0) &= \frac{1-Z}{1-\mathbb{E}(Z)}\left(Y-\mathbb{E}(Y|Z=0)\right) \\ \mathsf{EIF}(\alpha_1) &= \frac{Z}{\mathbb{E}(Z)}\left(A-\mathbb{E}(A|Z=0)\right) \\ \mathsf{EIF}(\alpha_0) &= \frac{1-Z}{1-\mathbb{E}(Z)}\left(A-\mathbb{E}(A|Z=0)\right). \end{aligned} \tag{13}$$

Finally, note that

$$\theta^*_{\mathsf{LATE}}(F_0)=\frac{\mu_1(F_0)-\mu_0(F_0)}{\alpha_1(F_0)-\alpha_0(F_0)},$$

so by the tricks in equation (10) and the results in equations (12) and (13), we obtain the EIF as equation (9).

4 Incorporating covariates

It is possible to incorporate covariates in to the IV framework. Let X the covariate. One of the simplest approach is to modify IV1-5 to the following form

1. **IV1= IV1x (exclusion restriction):** $Y(a,z)=Y(a,z')$ for all $a,z,z'\in\{0,1\}$.
2. **IV2x (unconfounded/randomization):** $Z\perp(Y(a,z),A(z'):a,z,z'\in\{0,1\})|X$ almost surely.
3. **IV3x (no defier/monotonicity):** Given X, $A(1)\geq A(0)$ almost surely.

4. **IV4x (IV relevance):** $\mathbb{E}(A|Z=1,X) \neq \mathbb{E}(A|Z=0,X)$ almost surely.

5. **IV5 (positivity):** $1 > P(Z=1|X) > 0$ almost surely.

Only IV1x stays the same as IV1. Other conditions are modified into 'given the covariates X'. All the derivations will hold in this case. The identified conditional LATE will be

$$\theta^*_{\mathsf{LATE}}(x) = \frac{\mathbb{E}(Y|Z=1,x) - \mathbb{E}(Y|Z=0,x)}{\mathbb{E}(A|Z=1,x) - \mathbb{E}(A|Z=0,x)}$$

and we have

$$\theta^*_{\mathsf{LATE}}(x) = \theta_{\mathsf{LATE}}(x) = \mathbb{E}(Y(1)-Y(0)|\mathsf{complier},x)$$

under assumption IV1-5x.

5 Linear structural equation model

An interesting interpretation of

$$\theta^*_{\mathsf{LATE}} = \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(A|Z=1) - \mathbb{E}(A|Z=0)}$$

is that this quantity can be expressed as the ratio of two slopes, i.e.,

$$\theta_{\mathsf{LATE}} = \frac{\beta_{Y\sim Z}}{\beta_{A\sim Z}},$$

where $\beta_{Y\sim Z}$ is the slope of fitting a linear model with response Y and covariate Z.

This insight implies a boarder view of IV under continuous A and Z and with the unmeasured confounders U under linear structural equation model (SEM). Consider the following linear SEM:

$$\begin{aligned}
&Z \sim p_Z \\
&U \sim p_U \\
&A = \alpha_A + \gamma Z + \eta_A U + \varepsilon_A \\
&Y = \alpha_Y + \beta A + \eta_Y U + \varepsilon_Y,
\end{aligned}$$

where $\varepsilon_A, \varepsilon_Y, Z, U$ are independent. The actual treatment effect is β.

If we fit a linear model with response A and covariate Z, the regression coefficient of Z will be $\beta_{A\sim Z} = \gamma$. If we fit a linear model with response Y and covariate Z, the regression coefficient of Z will be $\beta_{Y\sim Z} = \gamma\beta$. To see this, replacing A in the equation of Y leads to

$$Y = (\alpha_Y + \beta\alpha_A) + \beta\gamma Z + (\eta_Y + \beta\eta_A)U + (\varepsilon_Y + \beta\varepsilon_A).$$

Thus, the ratio will be

$$\frac{\beta_{Y\sim Z}}{\beta_{A\sim Z}} = \frac{\beta\gamma}{\gamma} = \beta,$$

which is the desired quantity.

6 Multiplicative local average treatment effect

In addition to the LATE, another treatment effect of interest is the multiplicative local average treatment effect (MLATE), which is defined as

$$\theta_{\mathsf{MLATE}} = \frac{\mathbb{E}(Y(1)|\mathsf{complier})}{\mathbb{E}(Y(0)|\mathsf{complier})}.$$

Under the same IV assumptions (IV1-5), one can show that this quantity is identified by the following quantity:

$$\theta_{\mathsf{MLATE}} = \frac{\mathbb{E}(YA|Z=1) - \mathbb{E}(YA|Z=0)}{\mathbb{E}(Y(1-A)|Z=1) - \mathbb{E}(Y(1-A)|Z=0)}.$$

One can also derive the EIF of MLATE using a similar derivation as Section 3. Also, one can incorporate the covariates into the MLATE and defined it as a conditional MLATE

$$\theta_{\mathsf{MLATE}}(x) = \frac{\mathbb{E}(Y(1)|\mathsf{complier}, X=x)}{\mathbb{E}(Y(0)|\mathsf{complier}, X=x)},$$

which can be identified by the following quantity

$$\theta_{\mathsf{MLATE}}(x) = \frac{\mathbb{E}(YA|Z=1, X=x) - \mathbb{E}(YA|Z=0, X=x)}{\mathbb{E}(Y(1-A)|Z=1, X=x) - \mathbb{E}(Y(1-A)|Z=0, X=x)}.$$

under IV1x-IV5x.