Lecture 6: Instrumental Variables

Tianpei Xie

Sep. 22nd., 2022

Contents

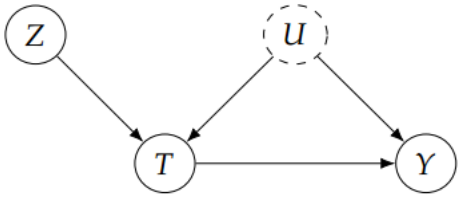

Figure 9.1: Graph where U is an unobserved confounder of the effect of T on Y and Z is an instrumental variable.

Figure 1: The causal graph for instrumental variable Z [Neal, 2020]

1 What is an Instrument?

- How can we identify causal effects when we are in the presence of unobserved confounding? One popular way is to find and use ***instrumental variables*** [Imbens and Rubin, 2015, Peters et al., 2017, Neal, 2020].
- **Instrumental variables (IVs)** is an alternative causal inference technique that does not rely on the ignoreability assumption. That is, it is used when there are ***unmeasured* confounders**.
- There are **three main assumptions** that must be satisfied for a variable to be considered an ***instrument***:

  1. **Assumption 1.1** ***(Relevance)***
     Z has a causal effect on ***treatment*** *T*
  2. **Assumption 1.2** ***(Exclusion Restriction)***
     Z causal effect on outcome Y is ***fully mediated*** *by treatment T*
  3. **Assumption 1.3** ***(Instrumental Unconfoundedness)***
     There are ***no backdoor paths*** *from Z to Y.*

- Intuitively, we interpret the instrumental variable Z as ***encouragement***, i.e. $Z = 1$ would *encourage* the treatment $T = 1$ on subject. To guarantee that the *Exclusion Restriction assumption* holds, Z **should not affect the unobserved confounder** U.
- This leads to an ***encouragement design***. Instead of performing randomized trial on *treatment* T, we can **randomize on the *encouragement*** $Z = 1$ vs. $Z = 0$. And then an ***intention-to-treat analysis*** would focus on *the causal effect of* ***encouragement***.

  On the other hand, we would still care about the causal effect of treatment itself, which would be the focus of this chapter.
- We may consider *treatment* ***assignment*** as an instrumental variable Z and the *treatment* ***received*** as the treatment variable T. Typically, not everyone assigned treatment will actually receive it (i.e. $Z \neq T$). This lead to ***randomized trials with non-compliance***. Non-compliance makes the randomized trials look like an observational study. The confounding is based on actual treatment received.
- For IVs, the **exclusion restriction / instrumental unconfoundedness** is a strong assumption.

2 Linear Setting

2.1 Binary Linear Setting

- In this section, we consider the following *noiseless linear causal structure models*

$$Y = \delta\, T + \alpha_u\, U \tag{1}$$

where U is the unobserved confounder. Note that Z does not appear in this equation due to exclusion restriction, since T is determined by Z.

- From (1), the causal effect of T on Y is defined as δ. To identify δ, we compute

$$\begin{aligned}
&\mathbb{E}\left[Y \mid Z=1\right] - \mathbb{E}\left[Y \mid Z=0\right] \\
&= \mathbb{E}\left[\delta\, T + \alpha_u\, U \mid Z=1\right] - \mathbb{E}\left[\delta\, T + \alpha_u\, U \mid Z=0\right] \\
&= \delta\left(\mathbb{E}\left[T \mid Z=1\right] - \mathbb{E}\left[T \mid Z=0\right]\right) + \alpha_u\left(\mathbb{E}\left[U \mid Z=1\right] - \mathbb{E}\left[U \mid Z=0\right]\right) \\
&(Z \perp\!\!\!\perp U \text{ by instrumental unconfoundedness}) \\
&= \delta\left(\mathbb{E}\left[T \mid Z=1\right] - \mathbb{E}\left[T \mid Z=0\right]\right) + \alpha_u\left(\mathbb{E}\left[U\right] - \mathbb{E}\left[U\right]\right) \\
&= \delta\left(\mathbb{E}\left[T \mid Z=1\right] - \mathbb{E}\left[T \mid Z=0\right]\right)
\end{aligned}$$

- **Proposition 2.1** *The causal effect of T on Y is identified by*

$$\delta = \frac{\mathbb{E}\left[Y \mid Z=1\right] - \mathbb{E}\left[Y \mid Z=0\right]}{\mathbb{E}\left[T \mid Z=1\right] - \mathbb{E}\left[T \mid Z=0\right]}. \tag{2}$$

The statistical quantity on the right is called ***Wald estimand****.*

- The ***Wald estimator*** is

$$\hat{\delta} = \frac{\sum_{i:z_i=1} y_i - \sum_{i:z_i=0} y_i}{\sum_{i:z_i=1} t_i - \sum_{i:z_i=0} t_i} \tag{3}$$

- An alternative explanation for (2) is from *causal graphical model* in Figure 1. For linear SCM, the causal association from T to Y is computed by the ***product of the coefficients*** along the directed path from T to Y. Due to the existance of unmeasured confounder U, the causal association is not directly measureable. Rather, we can measure total association, and unblocked backdoor paths also contribute to total association, which is why $\mathbb{E}\left[Y \mid T=1\right] - \mathbb{E}\left[Y \mid T=0\right] \neq \delta$. Because there are no backdoor paths from the instrument Z to Y, we can trivially identify the effect of Z on Y as $\mathbb{E}\left[Y \mid Z=1\right] - \mathbb{E}\left[Y \mid Z=0\right] = \alpha_z\delta$. Similarly, we can identify the effect of the instrument on $\mathbb{E}\left[T \mid Z=1\right] - \mathbb{E}\left[T \mid Z=0\right] = \alpha_z$. Then, we can divide the effect of Z on Y by the effect of the Z on T to identify $\delta = \frac{\alpha_z\delta}{\alpha_z}$.

2.2 Continuous Linear Setting

- We can generalize the setting to continuous Z and T, rather than binary.
- The Wald estimand for **continous** Z and T is

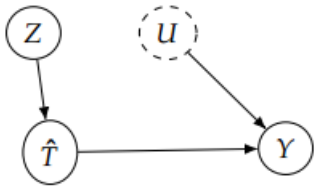

Figure 9.5: Augmented version of Figure 9.4, where T is replaced with $\hat{T} = \hat{\mathbb{E}}[T \mid Z]$, which doesn't depend on U, so there it no longer has an incoming edge from U.

Figure 2: The two-stage least square process. At first stage, $\widehat{T} = \mathbb{E}[T|\,|\,Z]$ replaces T, which does not depend on U [Neal, 2020]

Proposition 2.2

$$\delta = \frac{Cov(Y\,Z)}{Cov(T\,Z)}. \tag{4}$$

where $Cov(A\,B)$ is the cross-covariance between A and B.

Proof:

$$\begin{aligned}
\mathrm{Cov}(Y\,Z) &= \mathbb{E}[Y\,Z] - \mathbb{E}[Y]\;\mathbb{E}[Z] \\
&= \mathbb{E}[(\delta\,T + \alpha_u\,U)\;Z] - \mathbb{E}[\delta\,T + \alpha_u\,U]\;\mathbb{E}[Z] \\
&= \delta\;(\mathbb{E}[T\,Z] - \mathbb{E}[T]\,\mathbb{E}[Z]) + \alpha_u\;(\mathbb{E}[U\,Z] - \mathbb{E}[U]\,\mathbb{E}[Z]) \\
&\quad (Z \perp\!\!\!\perp U \text{ by instrumental unconfoundedness}) \\
&= \delta\;(\mathbb{E}[T\,Z] - \mathbb{E}[T]\,\mathbb{E}[Z]) + \alpha_u\;(\mathbb{E}[U]\,\mathbb{E}[Z] - \mathbb{E}[U]\,\mathbb{E}[Z]) \\
&= \delta\;(\mathbb{E}[T\,Z] - \mathbb{E}[T]\,\mathbb{E}[Z]) = \mathrm{Cov}(T\,Z) \qquad \blacksquare
\end{aligned}$$

- The ***Wald estimator*** for continous Z and T is

$$\hat{\delta} = \frac{\widehat{\mathrm{Cov}}(Y\,Z)}{\widehat{\mathrm{Cov}}(T\,Z)} \tag{5}$$

- An *equivalent estimator* to (5) is ***two-stage least square estimator***. The two stages are:

  1. **Linearly regress** T on Z to estimate $\mathbb{E}[T \mid Z]$ This gives us the projection $\widehat{T}$ of T onto Z;

     At this stage, T is replaced by $\widehat{T} = \mathbb{E}[T \mid Z]$ which does not depend on U. Because $\widehat{T}$ isn't a function of U, we can think of removing the $U \to \widehat{T}$ edge in this graph. See Figure 2.

  2. **Linearly regress** Y on $\widehat{T}$ to estimate $\mathbb{E}[Y \mid \widehat{T}]$. Obtain our estimate $\hat{\delta}$ as the fitted coefficient in front of $\widehat{T}$.

     Because there are no backdoor paths from $\widehat{T}$ to Y, we can get that association is causation in stage two, where we simply regress Y on $\widehat{T}$ to estimate the causal effect.

3 Nonparametric Identification of Local ATE

The problem with approaches above is that we made explicit assumption on the ***parameteric form*** of causal effect, i.e. the ***linearity*** assumption. For example, this assumption requires ***homogeneity*** (that the *treatment effect is the same for every unit*). There are other variants that encode the homogeneity assumption. Ideally, wed be able to use instrumental variables for identification without making any parametric assumptions such as linearity or homogeneity.

3.1 New Potential Notation with Instruments

- Define the ***potential treatment value*** $T(1) := T(Z = 1) = T^{do(Z=1)}$ and $T(0) := T(Z = 0) = T^{do(Z=0)}$ as the treatment we would take if we were to get instrument value $Z = 1$ and $Z = 0$, respectively.
- We can define the **average causal effect of treatment *assignment* on treatment *received*** as

$$\mathbb{E}\left[T(1) - T(0)\right] = \mathbb{E}\left[T^{do(Z=1)}\right] - \mathbb{E}\left[T^{do(Z=0)}\right] \tag{6}$$

  This is **the propotion treated** if ***everyone*** has been assigned to receive the treatment, ***minus*** the propotion treated if ***no one*** has been assigned to receive the treatment. If **perfectly compliance**, this quantity is equal to 1.

- We also consider the **average causal effect of treatment *assignment* on *outcome***

$$\mathbb{E}\left[Y(Z = 1) - Y(Z = 0)\right] = \mathbb{E}\left[Y^{do(Z=1)}\right] - \mathbb{E}\left[Y^{do(Z=0)}\right] \tag{7}$$

  This is the **average values of outcome** if ***everyone*** has been assigned to receive the treatment, ***minus*** the average values of outcome if ***no one*** has been assigned to receive the treatment. This is also called the ***intention-to-treat effect (ITT)***. If perfectly compliance, this quantity is equal to the average causal effect (ATE).

- Due to ***instrumental unconfoundedness*** and consistency, we can estimate $\mathbb{E}\left[T(1)\right] = \mathbb{E}\left[T \mid Z = 1\right]$ and $\mathbb{E}\left[T(0)\right] = \mathbb{E}\left[T \mid Z = 0\right]$. Similarly, we can estimate $\mathbb{E}\left[Y(Z = 1)\right] = \mathbb{E}\left[Y \mid Z = 1\right]$ and $\mathbb{E}\left[Y(Z = 0)\right] = \mathbb{E}\left[Y \mid Z = 0\right]$.

3.2 Principal Stratification

- **Definition** (***Principal Strata***) [Imbens and Rubin, 2015, Neal, 2020]
  1. ***Compliers*** - always take the treatment that they are *encouraged* to take. Namely, $T(1) = 1$ and $T(0) = 0$. This is the targeted sub-population.
  2. ***Always-takers*** - always take the treatment, **regardless** of encouragement. Namely, $T(1) = 1$ and $T(0) = 1$.
  3. ***Never-takers*** - never take the treatment, **regardless** of encouragement. Namely, $T(1) = 0$ and $T(0) = 0$.

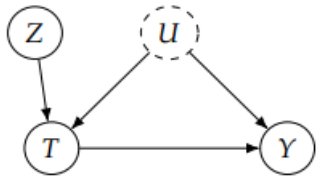

Figure 9.6: Causal graph for the compliers and defiers.

(a)

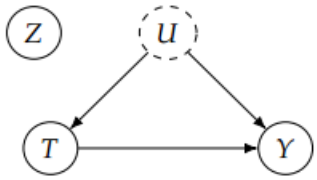

Figure 9.7: Causal graph for the always-takers and never-takers.

(b)

Figure 3: The causal graph for (a) complier and defier; (b) always-taker and never taker

4. ***Defiers*** - always take the **opposite treatment** of the treatment that they are *encouraged* to take. Namely, $T(1) = 0$ and $T(0) = 1$.

- The principal strata is a partition of subjects into four sub-populations.
  - Clearly, the encouragement will not work for **Never-takers**. We will ***not*** *learn* ***anything*** *on causal effect* of treatment from people in this subpopluation, $P(T = 1|\boldsymbol{x}) = 0$.
  - For **Compliers**, they receive treatments when encouraged to, and do not otherwise. The treatment received in this subpopulation is ***randomized***.
  - For **Defiers**, their treatment received is also ***randomized*** but in the *opposite* way. Usually we would image them not existed since normally the group who are not assigned to the treatment has no access to it.
  - For **Always-takers**, there is **no information on causal effect** from this subpopulation either since no variations in the treatment received $P(T = 0|\boldsymbol{x}) = 0$.
- Given some **observed** value of Z and T, **we cant actually identify which stratum were in**, since we cannot observe the treatment received under the counterfactual treatment assignment. There are four combinations of the binary variables Z and T; for each of these combinations, well note that *more than one stratum* is ***compatible*** with the observed combinations of values.
  1. Given $Z = 1$, $T = 1$, the compatible strata: ***compliers*** or ***always-takers***;
  2. Given $Z = 1$, $T = 0$, the compatible strata: ***defiers*** or ***never-takers***;
  3. Given $Z = 0$, $T = 1$, the compatible strata: ***defiers*** or ***always-takers***;
  4. Given $Z = 0$, $T = 0$, the compatible strata: ***compliers*** or ***never-takers***.

3.3 Local ATE

- A main motivation for using instrument variables is that there exists unobserved confounders. In this situation, the covariate adjustment does not work since we cannot marginalize over ***all*** confounders. This means that **we cannot identify the average treatment effect for the entire population**.
- IV methods focus on the identification of ***local average treatment effect (LATE)***. It is

also known as ***complier average treatment effect (CATE)***.

- **Definition** ***(Local Average Treatment Effect (LATE) / Complier Average Causal Effect (CACE))*** [Imbens and Rubin, 2015, Neal, 2020]

$$\mathbb{E}\left[Y(T=1) - Y(T=0) \,|\, T(Z=1)=1,\, T(Z=0)=0\right] = \mathbb{E}\left[Y(1) - Y(0) \,|\, \text{compliers}\right] \quad (8)$$

It is the average treatment effect conditioned on the ***complier sub-populations***.

This is a *causal quantity* since it ***contrasts counterfactuals*** in a ***common*** sub-populations. Note that this quantity make no inference on other populations due to no interference assumption.

- Given observed Z and T, the **challenge** is that we cannot identify the complier directly.
- To identify the LATE, although we will need to introduce a new assumption.

Assumption 3.1 ***(Monotonicity)***

$$\forall i, \quad T_i(Z=1) \geq T_i(Z=0) \quad (9)$$

This is equivalent to say that ***there is no defier****. That is, if we are encouraged to take the treatment ($Z=1$), we are either more likely or equally likely to take the treatment than we would be if we were encouraged to not take the treatment ($Z=0$).*

- Under monotonicity assumption,
  1. Given $Z=1$, $T=1$, the compatible strata: ***compliers*** or ***always-takers***;
  2. Given $Z=1$, $T=0$, the compatible strata: ***never-takers***;
  3. Given $Z=0$, $T=1$, the compatible strata: ***always-takers***;
  4. Given $Z=0$, $T=0$, the compatible strata: ***compliers*** or ***never-takers***.
- Under monotonicity assumption, we can derive the *nonparametric identification* result for the LATE estimand:

Theorem 3.2 ***(LATE Nonparametric Identification)*** *[Neal, 2020]*
Given that Z is an instrument, Z and T are binary variables, and that monotonicity holds, the following is true:

$$\mathbb{E}\left[Y(1) - Y(0) \,|\, T(1)=1,\, T(0)=0\right] = \frac{\mathbb{E}\left[Y \,|\, Z=1\right] - \mathbb{E}\left[Y \,|\, Z=0\right]}{\mathbb{E}\left[T \,|\, Z=1\right] - \mathbb{E}\left[T \,|\, Z=0\right]} \quad (10)$$

Proof: Well start with the causal effect of Z on Y and decompose it into weighted stratum-specific causal effects using the law of total probability:

$$\mathbb{E}\left[Y^{do(Z=1)} - Y^{do(Z=0)}\right]$$
$$= \sum_{i\in\{0,1\}, j\in\{0,1\}} \mathbb{E}\left[Y^{do(Z=1)} - Y^{do(Z=0)} \,\middle|\, T^{do(Z=1)}=i,\, T^{do(Z=0)}=j\right] P(T^{do(Z=1)}=i,\, T^{do(Z=0)}=j)$$

Note that by monotonicity assumption, $P(T^{do(Z=1)}=0,\, T^{do(Z=0)}=1)=0$. Morover, the causal effect of always-takers and never-takers are zero. $\mathbb{E}\left[Y|Z, \text{always-takers}\right] = \mathbb{E}\left[Y|\text{always-takers}\right]$ and $\mathbb{E}\left[Y|Z, \text{never-takers}\right] = \mathbb{E}\left[Y|\text{never-takers}\right]$. So the LATE is only on complier subset

$$\mathbb{E}\left[Y^{do(Z=1)} - Y^{do(Z=0)}\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[Y^{do(Z=1)} - Y^{do(Z=0)} \,\middle|\, T^{do(Z=1)} = 1,\, T^{do(Z=0)} = 0\right] P(T^{do(Z=1)} = 1,\, T^{do(Z=0)} = 0) \\
\Rightarrow \quad &\mathbb{E}\left[Y^{do(Z=1)} - Y^{do(Z=0)} \,\middle|\, T^{do(Z=1)} = 1,\, T^{do(Z=0)} = 0\right] \\
&= \frac{\mathbb{E}\left[Y^{do(Z=1)} - Y^{do(Z=0)}\right]}{P(T^{do(Z=1)} = 1,\, T^{do(Z=0)} = 0)} \qquad (11)
\end{aligned}
$$

For complier, the treatment received will equal to treatment assigned so $Y^{do(Z=i)} = Y^{do(T=i)}$ for $i \in \{0, 1\}$. So we can rewrite the LHS of (11) as

$$
\begin{aligned}
\mathbb{E}\left[Y(1) - Y(0) \,\middle|\, T(1) = 1,\, T(0) = 0\right] &= \frac{\mathbb{E}\left[Y(Z=1)\right] - \mathbb{E}\left[Y(Z=0)\right]}{P(T(1) = 1,\, T(0) = 0)} \qquad (12) \\
&\quad \text{by instrumental unconfoundedness} \\
&= \frac{\mathbb{E}\left[Y \mid Z = 1\right] - \mathbb{E}\left[Y \mid Z = 0\right]}{P(T(1) = 1,\, T(0) = 0)} \qquad (13)
\end{aligned}
$$

Note that we cannot identify the complier from observation alone, but we can identify the never-takers $(T = 0 \mid Z = 1)$ and always-takers $(T = 1 \mid Z = 0)$ due to monotonicity assumption. So

$$
\begin{aligned}
P(T(1) = 1,\, T(0) = 0) &= 1 - P(T(1) = 0,\, T(0) = 0) - P(T(1) = 1,\, T(0) = 1) \\
&= 1 - P(T = 0 \mid Z = 1) - P(T = 1 \mid Z = 0) \\
&= P(T = 1 \mid Z = 1) - P(T = 1 \mid Z = 0) \qquad (14) \\
&= P\left(\text{always-taker } \textit{OR} \text{ complier}\right) - P\left(\text{always-taker}\right) > 0
\end{aligned}
$$

So substituting (14)

$$
\begin{aligned}
\mathbb{E}\left[Y(1) - Y(0) \,\middle|\, T(1) = 1,\, T(0) = 0\right] &= \frac{\mathbb{E}\left[Y \mid Z = 1\right] - \mathbb{E}\left[Y \mid Z = 0\right]}{P(T = 1 \mid Z = 1) - P(T = 1 \mid Z = 0)} \qquad (15) \\
&= \frac{\mathbb{E}\left[Y \mid Z = 1\right] - \mathbb{E}\left[Y \mid Z = 0\right]}{\mathbb{E}\left[T \mid Z = 1\right] - \mathbb{E}\left[T \mid Z = 0\right]}
\end{aligned}
$$

The last equality holds since T is binary variable. ∎

- Same as result under linear setting, the LATE estimand is identified by **Wald estimand**:

$$
\text{LATE} = \delta = \frac{\mathbb{E}\left[Y \mid Z = 1\right] - \mathbb{E}\left[Y \mid Z = 0\right]}{\mathbb{E}\left[T \mid Z = 1\right] - \mathbb{E}\left[T \mid Z = 0\right]} = \frac{\text{Average Causal Effect (ITT) } (Z \to Y)}{\text{Average Causal Effect } (Z \to T)}.
$$

Different from (2), the Wald estimand is used to compute the *local* ATE. Because there are *no backdoor paths* from the instrument Z to Y. Both the ***average causal effect of treatment assignment on outcome (ITT)*** and the ***average treatment effect of treatment assignment on treatment received*** are **identifiable**.

- If perfect complier, LATE = ITT; otherwise LATE ≥ ITT.

  ITT is an ***under-estimate*** of the LATE since there are some people who were assigned to the treatment but did not take it.

4 Instrumental variables in observational studies

- The instrumental variables can be thought of a randomizer in the natural experiment.
- The key challenge is to think of a variable that affects the treatment but does not affect the outcome *directly*. Note that only assumptions that affect the treatment can be checked with observations.
- The validity of exclusion restriction and instrumental unconfoundedness largely depends on the domain expertise.
- For example, the ***calandar time*** can be used as an IV, e.g. the treatment probability change over time and Z can be early time period vs. late time period. In order to make sure exclusion restriction, the other treatment practices and patient behaviors should not change during two different time periods.
- Also, the ***distance*** to a specialty care center is used as an IV for health outcomes.
- The health care provider preference, i.e. the treatment prescribed to previous patients, can be used as an IV.
- ***Mendelian randomization***: some genetic variant is associated with some behavior (e.g. alcohol use), but is assumed to not be associated with the outcome.
- In these observational settings, we can still think about ***compliance***, i.e compliance with encouragement.
- In practice, it is always helpful to perform **sensitivity analysis** on the exclusion restriction assumption ("what would change if Z *directly affect* Y by a given amount ρ?") and the monotonicity assumption ("what would change if the proportion of defier is π?").

4.1 Weak instruments

- The strength of an IV is how well it **predicts** the treatment received.
  - A ***strong*** *instrument* is highly predictive of the treatment received. That is, the encouragement will greatly increase the probability of treatment;
  - A ***weak*** *instrument* is weakly predictive of the treatment received. That is, the encouragement will barely increase the probability of treatment.
- We can measure the ***strength of instrumental variable*** using the ***probability of compliers***

$$\mathbb{E}\left[T \mid Z=1\right] - \mathbb{E}\left[T \mid Z=0\right] \tag{16}$$

- A weak instrument can cause problems. Suppose only 1% of population are compliers, the amount of samples is $1\%\, n$ for a total of n samples. The **limited sample size** would cause **high variance** for estimators. The causal effect estimator is unstable.
- If IV is weak, an IV analysis may not be the best option. There are researches to *strengthen* the IV (using ***near/far matching*** for example).

References

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Brady Neal. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*, 2020.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.