DS102 - Discussion 7

October 14, 2020

1 Background on instrumental variables

In our lectures on causal inference, we've looked at two extremes on how to infer treatment effects: the *randomized trial* where we have complete control of the treatments, and the observational study where we have no control over the treatments, but try to estimate treatment effects from data we observe. *Instrumental variables* are a strategy that falls somewhere in between.

Suppose we are interested in determining whether reading more books causes students' SAT test scores to improve. It's not always a good idea to conduct a randomized trial, since we may not ethically or practically be able to force people to read or not read. On the other hand, if we were to just look at observational data, there might always be an unobserved confounding variable that interferes with our ability to infer the causal effect of reading. For example, one confounder might be a student's family's income, since it changes the educational resources (including both reading material and standardized test preparation) a student had growing up.

As something in between those two approaches, we might employ *encouragement design*. In this setting, we randomly select people and encourage them to read by organizing a "readathon" at their school. This encouragement, which we call an *instrumental variable* (IV), needs to satisfy two properties in order to use the method we develop today:

1. It has a causal effect on the treatment variable (here, how much a student reads).
2. It has no *direct* effect on the outcome variable (here, a student's SAT score), only indirectly through the treatment variable. (This condition also implies the IV has no effect on the confounder.)

Organizing a readathon has no effect on a student's SAT score directly or on a student's household income, but has a causal effect on the number of books a student will read.

Our encouragement design results in a dataset of n students, with the following variables:

- $Y^{(i)}$ is the SAT score of the the i-th student.
- $X_1^{(i)}$ is how many books the i-th student read over the last month.
- $Z^{(i)} \in \{0, 1\}$ indicates whether or not a readathon was organized at the i-th student's school (instrumental variable).

Finally, let $X_2^{(i)}$ denote the i-th student's family's income (a confounder), which we do not observe.

In the following problems, we will develop a method for using Z to estimate the causal effect of X_1 on Y, even though we know they are confounded by X_2. See Figure 1 for the causal graphical model of this setup.

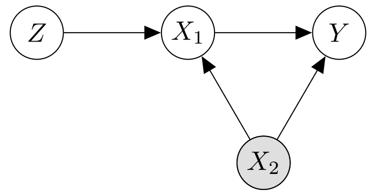

Figure 1: A causal graphical model showing the instrumental variable setup. The instrumental variable Z does not affect the confounder X_2 (income), nor does it affect the outcome Y (SAT score), except through the treatment X_1 (number of books read).

2 Instrumental variables and two-stage least squares (2SLS)

How can we use an instrumental variable Z to infer the causal effect of X_1 on Y? One approach is to model the causal relationship between Y and X_1 as a linear regression problem.

Let's assume that the i-th student's SAT score is generated through the following linear model:

$$Y^{(i)} = \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \epsilon^{(i)},$$

where β_1, β_2 are unknown coefficients, and $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$ is noise.

Our goal is to accurately estimate β_1, which tells us how $Y^{(i)}$ varies with $X_1^{(i)}$.

1. Before we incorporate the instrumental variable, let's first see what can go wrong when we *don't* employ encouragement design and include instrumental variables in the linear regression problem.

   Suppose we observe $X_1^{(i)}$ but not the confounding variable $X_2^{(i)}$. We decide to run a linear regression model on the observed variable X_1 only. Define the vectors

$$X_1 = \begin{bmatrix} X_1^{(1)} \\ X_1^{(2)} \\ \vdots \\ X_1^{(n)} \end{bmatrix}; \quad X_2 = \begin{bmatrix} X_2^{(1)} \\ X_2^{(2)} \\ \vdots \\ X_2^{(n)} \end{bmatrix}; \quad Y = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \\ \vdots \\ Y^{(n)} \end{bmatrix}.$$

   Find the least square estimator of β_1 which minimizes the term $\|Y - X_1\beta_1\|_2^2 = \sum_{i=1}^n (Y^{(i)} - X_1^{(i)}\beta_1)^2$, i.e. compute $\hat{\beta}_1 = \operatorname{argmin}_{\beta_1} \|Y - X_1\beta_1\|_2^2$. This is also known as ordinary least square (OLS) estimator.

   Solution: The solution can be derived via multiple ways. One can directly take the derivative with respect to β_1, let $F(\beta_1) = \|Y - X_1\beta_1\|_2^2 = Y^\top Y - 2Y^\top X_1\beta_1 + \beta_1^\top X_1^\top X_1 \beta$. Note that since β_1 is a scalar, it is a quadratic function. We can directly take the derivative and set it to be 0, which gives

$$\hat{\beta}_1 = (X_1^\top X_1)^{-1} X_1^\top Y.$$

As a more general solution of OLS estimator, if β_1 is a $d \times 1$ vector and X_1 is a $n \times d$ matrix, we can compute the derivative with respect to the vector as (as a reference, The matrix cook book provides a thorough list of the related vector and matrix computation.)

$$\frac{\partial F}{\partial \beta_1} = 2X_1^\top X_1 \beta - 2X_1^\top Y.$$

By setting the derivative to be 0, we have

$$\hat{\beta}_1 = (X_1^\top X_1)^{-1} X_1^\top Y.$$

2. Assuming that $n = 1$ for simplicity, can you think of a plausible situation where $\hat{\beta}_1$ is a biased estimator?

Solution: Suppose $n = 1$. A simple situation would be a case where X_1 depends on X_2. For example, we could have

$$\begin{aligned} X_1 &= \frac{X_2}{50000} + \epsilon' \\ \implies X_2 &= 50000X_1 + \epsilon'', \end{aligned}$$

where ϵ' and ϵ'' are both zero mean noise variables. This corresponds to the situation where a student is generally more likely to read more books if their family is wealthy.

$$\begin{aligned} \mathbb{E}\left[\hat{\beta}_1\right] &= \mathbb{E}\left[\frac{Y^{(1)}}{X_1^{(1)}}\right] \\ &= \mathbb{E}\left[\frac{\beta_1 X_1^{(1)} + \beta_2 X_2^{(1)} + \epsilon}{X_1^{(1)}}\right] \\ &= \beta_1 + 50000\beta_2. \end{aligned}$$

Depending on β_2, this can be an extremely biased estimator.

3. Now suppose we employ encouragement design: we incentivize a randomly chosen subset of students to read more books by organizing a readathon at their school. Let

$$Z = \begin{bmatrix} Z^{(1)} \\ Z^{(2)} \\ \vdots \\ Z^{(n)} \end{bmatrix},$$

where $Z^{(i)} = 1$ if the i-th student's school had a readathon and $Z^{(i)} = 0$ otherwise.

In this problem, we use our intuition to develop an estimator of the effect of X_1 on Y. Informally, we can think of β_1 as the rate of change of $Y^{(i)}$ with respect to $X_1^{(i)}$. Then it follows from the chain rule that

$$\frac{dY^{(i)}}{dX_1^{(i)}} = \frac{dY^{(i)}/dZ^{(i)}}{dX_1^{(i)}/dZ^{(i)}}.$$

An intuitive estimator of β_1 is then to estimate both the denominator and numerator of this fraction:

$$\hat{\beta}_{IV} = \frac{(Z^\top Z)^{-1} Z^\top Y}{(Z^\top Z)^{-1} Z^\top X_1}.$$

Show that

$$\hat{\beta}_{IV} = (Z^\top X_1)^{-1} Z^\top Y.$$

Solution: The solution can be seen by noting that $Z^\top Z$, $Z^\top Y$, $Z^\top X_1$ are scalar. Thus the $(Z^\top Z)^{-1}$ in both numerator and denominator cancel each other. As a more general case, if both Z and X_1 are $n \times d$ matrices, we have

$$\begin{aligned}
\hat{\beta}_{IV} &= ((Z^\top Z)^{-1} Z^\top X_1)^{-1} (Z^\top Z)^{-1} Z^\top Y \\
&= (Z^\top X_1)^{-1} (Z^\top Z)(Z^\top Z)^{-1} Z^\top Y \\
&= (Z^\top X_1)^{-1} Z^\top Y.
\end{aligned}$$

The purpose of simplifying this expression will become clear in Problem 4, when we compare it to the 2SLS procedure.

4. To formalize the estimator we derived in the previous problem, we now consider the *two-stage least squares estimator (2SLS)*. This estimator uses the instrumental variable Z to get a better estimate of the relationship β_1 between X_1 and Y, and has two stages:

   1. Find the OLS estimate with X_1 as the output and Z as the input:

   $$\hat{\alpha} = (Z^\top Z)^{-1} Z^\top X_1.$$

   2. Find the OLS estimate with Y as the output and $\hat{X}_1 = Z\hat{\alpha}$ as the input:

   $$\hat{\beta}_{2SLS} = (\hat{X}_1^\top \hat{X}_1)^{-1} \hat{X}_1^\top Y.$$

   Show that $\hat{\beta}_{2SLS} = \hat{\beta}_{IV}$.

Solution:

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= (Z^\top \hat{\alpha} Z \hat{\alpha})^{-1} Z^\top \hat{\alpha} Y \text{ (replacing } \hat{X}_1 \text{ with } Z\hat{\alpha}) \\
&= \hat{\alpha}^{-1} (Z^\top Z)^{-1} Z^\top Y \\
&= (Z^\top X_1)^{-1} (Z^\top Z)(Z^\top Z)^{-1} Z^\top Y \\
&= \hat{\beta}_{IV}
\end{aligned}
$$

This shows that the 2SLS procedure produces the same estimator as the one derived in Problem 2, where we used our intuition to combine the relationships between Y and Z and between X_1 and Z.