Fundamentals of Data Science

Instrumental variables

Ramesh Johari

Omitted variable bias

Omitting variables

Recall that a key problem in causal inference is *omitted variable bias*.

Suppose, for example, that we observe that individuals who are on a particular diet have lower rates of inflammatory arthritis.

However, we don't realize that these same individuals also have high levels of exercise.

If we ignore or do not observe exercise levels, we may mistakenly attribute the lower rates of inflammatory arthritis to the diet.

A simple model

Consider the following simple population model:

- There are two covariates X and Z.
- Given values of X and Z, the outcome Y is realized as:

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \varepsilon,$$

  where $\mathbb{E}[\varepsilon|X, Z] = 0$.
- We observe n i.i.d. observations from this model.

Omitting X

Suppose that we *don't observe* Z_i for each i.

We run the regression $Y_i \approx \hat{\beta}_0 + \hat{\beta}_X X_i$ using OLS.

What happens?

Omitting X

Recall that for a simple linear regression:

$$\hat{\beta}_X = \frac{\sum_{i=1}^n X_i Y_i - \overline{XY}}{\sum_{i=1}^n X_i^2 - \overline{X}^2}.$$

The numerator is the sample covariance of $\mathbf{X}$ and $\mathbf{Y}$, and the denominator is the sample variance of $\mathbf{X}$.

We now show that this is a *biased* estimator of β_X.

Omitted variable bias

In particular we can show:

$$\mathbb{E}[\hat{\beta}_X | \mathbf{X}, \mathbf{Z}] = \beta_X + \beta_Z \hat{\delta}_{Z|X},$$

where $\hat{\delta}_{Z|X}$ is the coefficient of X in the regression $Z_i \approx \hat{\delta}_0 + \hat{\delta}_1 X_i$:

$$\hat{\delta}_{Z|X} = \frac{\sum_{i=1}^n X_i Z_i - \overline{XZ}}{\sum_{i=1}^n X_i^2 - \overline{X}^2}.$$

The idea is that by omitting Z, we estimate a coefficient on X that includes not only the direct effect of X, but also a part of the effect of Z (determined exactly by how correlated Z is with X).

Correcting omitted variable bias

How can we correct for omitted variable bias?

- We can run a randomized experiment where we directly vary X; e.g., we can set one value of X in the control group, and another value of X in the treatment group. This forces X to be independent of Z in the population, eliminating the bias.
- We can add additional covariates, in hopes that this removes any bias in our estimates.
- These strategies are limited in their applicability. *Instrumental variables* are another approach to addressing this problem.

Instrumental variables

What is an instrument?

Informally, an *instrument* is an additional covariate W that has two properties:

- *"Strong first stage"*: The instrument is *positively correlated* with the covariate X. In other words, when W varies, X varies as well.
- *"Exclusion restriction"*: Given X, the instrument is *uncorrelated* with the outcome Y.

The first property allows the instrument to act as a "knob" that adjusts X. This can be directly verified from the data.

The second property ensures that any effects observed from this knob are only felt because of its action through X, rather than through any direct effect on the outcome. A convincing case must be made for this property based on the structure of the setting; usually data analysis is not enough.

Example: Education and earnings

A classic example is found in Angrist and Krueger's study of the effect of compulsory schooling on earnings.

In studying the effect of schooling on earnings is that there is an omitted variable bias: econometricians cannot directly observe *individual ability*.

Angrist and Krueger develop a clever instrument: they note that due to features of regulations around compulsory schooling, students born earlier in the year can drop out of school with less schooling than students born later in the year.

This allows them to use *quarter of birth* as an instrument for *quantity of education*.

Example: Supply shocks

Suppose you want to measure the sensitivity of price a product to the supply of a given product.

This is a challenging problem because supply is *endogenous*: low prices encourage greater supply.

However, in some contexts, external events can alter the level of supply available, without directly influencing prices. For example, variations in weather can influence production of vegetable crops, allowing weather to be used as an instrument for supply.

From Dubner and Levitt (NY Times Magazine): "From an economist's perspective, the great thing about the weather is that there is nothing humans can do to affect it (at least until recently)."

Example: Encouragement design

Suppose you launch a new product, and want to measure satisfaction from use. It may be difficult (or unethical) to do a randomized experiment where you force half the population to use it, and exclude half the population from using it.

Instead, you can *randomize encouragement* to use the product: encourage half the population to use the product, and do not encourage the other half.

As long as:

1. encouragement is strongly correlated with usage; and
2. the act of encouragement itself is uncorrelated with satisfaction with the product,

encouragement acts as an instrument for usage.

An instrument W

Suppose in our example that we find a covariate W such that:

1. The sample covariance of $\mathbf{X}$ and $\mathbf{W}$ is positive in our data.
2. In the population, $\mathbb{E}[W\tilde{\varepsilon}|X] = 0$, where $\tilde{\varepsilon} = \beta_Z Z + \varepsilon$ is the total error if Z is an omitted variable. This is the exclusion restriction, and cannot be verified from data.

How do we use the instrument?

Estimating β_X

We want to estimate the causal effect of X on Y.

Intuitively, what do we expect?

- Suppose that a one unit change in W is associated with a δ unit change in X.
- Suppose in addition that a one unit change in W is associated with a γ unit change in Y.

Since variation in W is associated only to variation in X, the effect on Y could not have arisen from any other mechanism except variation in X. Thus we expect the effect of X on Y to be γ/δ.

Estimating β_X

Formally, we first regress X on W via $X_i \approx \hat{\delta}_0 + \hat{\delta} W_i$ to obtain:

$$\hat{\delta} = \frac{\sum_{i=1}^n X_i W_i - \overline{XW}}{\sum_{i=1}^n W_i^2 - \overline{W}^2}.$$

We also regress Y on W via $Y_i \approx \hat{\gamma}_0 + \hat{\gamma} W_i$ to obtain:

$$\hat{\gamma} = \frac{\sum_{i=1}^n Y_i W_i - \overline{YW}}{\sum_{i=1}^n W_i^2 - \overline{W}^2}.$$

Estimating β_X

Thus we have:

$$\frac{\hat{\gamma}}{\hat{\delta}} = \frac{\sum_{i=1}^{n} Y_i W_i - \overline{YW}}{\sum_{i=1}^{n} X_i W_i - \overline{XW}}.$$

It can be shown that:

$$\mathbb{E}\left[\frac{\hat{\gamma}}{\hat{\delta}}\right] = \beta_X.$$

In other words, this is an *unbiased* estimator of the true causal effect of X on Y.
This is called the *instrumental variables least squares* estimator.

Two-stage least squares

Computing estimates via instrumental variables

Suppose there are many instruments: a matrix $\mathbf{W}$, where W_{ik} is the i'th observation of the k'th instrument.

Suppose these are instruments for covariates in the matrix $\mathbf{X}$, where X_{ij} is the i'th observation of the j'th covariate (sometimes referred to as the *endogenous* covariates).

Suppose in addition we have additional covariates $\mathbf{U}$, where $U_{i\ell}$ is the i'th observation of the ℓ'th covariate (sometimes referred to as the *exogenous* covariates).

And finally suppose we have outcomes Y_i.

Computing estimates via instrumental variables

In practice, with many instruments and many covariates, the technique used to construct an estimator of the effect of the covariates in $\mathbf{X}$ is referred to as *two-stage least squares* (2SLS). The basic idea of 2SLS is as follows:

1. Regress each covariate in $\mathbf{X}$ on $\mathbf{W}$ and all the other covariates in $\mathbf{U}$. Use this to get fitted values $\hat{X}_{ip}$ for each covariate.
2. Regress the outcomes $\mathbf{Y}$ on the fitted values $\hat{\mathbf{X}}$ and the covariates in $\mathbf{U}$.

The idea is that the variation in the instruments "explains" variation in the covariates in $\mathbf{X}$. The exclusion restriction assumes that any variation in the outcome can only have arisen due to this variation in the covariates in $\mathbf{X}$, not due to variation in the instruments.

2SLS: Things to know

- In the simple case we've discussed thus far, 2SLS gives the same answer as our earlier computation of the IV least squares estimator.
- Despite being called "two-stage", 2SLS is not estimated in two stages: it is solved simultaneously.
- This is important especially for computation of standard errors, which would be incorrect if we just reported second stage standard errors.

Concluding thoughts

The challenge of finding instruments

Instrumental variables are typically found when there is exogenous variation that leads to changes in the variable of interest to the data scientist.

It can be hard to verify the exclusion restriction, especially if all one has is observational data.

However, variation induced by experiments can often be used as an instrumental variable. This is a common use case in instrumental variables analysis in, e.g., the tech industry.