Causal Inference with Observational Data

3. Instrumental Variables

Marc F. Bellemare

May 2018

Introduction

Having discussed the importance of making causal statements, the remainder of these lecture notes will be dedicated to the methods favored by applied microeconomists nowadays when trying to make causal statements.

I say "nowadays," because economics—applied microeconomics, at least—is not immune to fads and fashions.

Let us start with instrumental variables (IV), which is perhaps the oldest method used by economists to tease out causal relationships from messy observational (i.e., nonexperimental) data.

Introduction

The setup for IV estimation is pretty simple. We start from the equation

$$y_i = \alpha + \beta x_i + \gamma D_i + \epsilon_i, \tag{1}$$

which is similar to the core equation laid out in the previous chapter, except that the variable x now denotes a vector of control variables and the variable D our variable of interest, i.e., the variable whose causal effect on y we are interested in estimating.

We will soon see that the vector of control variables x can be very important when doing IV.

Introduction

Typically, estimating equation 1 does not yield a causal estimate of γ, only a partial correlation coefficient. What IV does is to condition D on an IV (or a vector of IVs) z (let us set aside what requirements z should satisfy for the time being), such that

$$D_i = \theta + \lambda x_i + \pi z_i + \xi_i. \tag{2}$$

From equation 2, we obtain $\widehat{D}_i$, which is such that

$$\widehat{D}_i = \widehat{\theta} + \widehat{\lambda} x_i + \widehat{\pi} z_i. \tag{3}$$

Introduction

We then substitute $\widehat{D}_i$ for D_i in equation 1 such that, with a slight abuse of notation, we estimate

$$y_i = \alpha + \beta x_i + \gamma \widehat{D}_i + \epsilon_i. \tag{4}$$

Let us set aside for a second how we estimate equations 2 and 4 (i.e., 2SLS or simultaneously) in order to focus on the intuition behind the setup just laid out.

Introduction

Before discussing that intuition, we must discuss the requirements imposed on z for it to be a good IV. There are two such requirements:

1. *Relevance*. The coefficient π on z in equation 2 has to be significantly different from zero. What is more, the F-statistic for π (i.e., the square of the t-statistic for the same coefficient) has to be above 13 or so in order for your IV not to be a weak IV. This is testable.
2. *Exclusion Restriction*. The variable z has to affect y only through D. This is not testable.

We will get back to those requirements—the second in particular—after discussing the intuition behind IV.

Introduction

Intuitively, what IV does is to purge D of its correlation with ϵ—that is, it gets rid of the "bad" variation in D, which compromises identification because it entails that $Cov(D, \epsilon) \neq 0$—while keeping only the "good" variation in D—that part of the variation in D that is uncorrelated with ϵ.

That is why the exclusion restriction is so important: Without it, D still retains some of that bad variation, your estimate of γ is not identified, and the estimated relationship cannot be argued to be causal.

Introduction

But even when your estimate of γ is identified and can be argued to be causal, the problem is that γ_{IV} is not comparable to γ_{OLS}—the latter is an average treatment effect (ATE), the former is a local average treatment effect (LATE).

Intuitively, the LATE is the ATE for those units of observation that were induced to take up treatment D as a result of the IV. We'll get back to this in a few minutes.

What Is a Good IV?

Good IVs are hard to come by, so much so that very often, a research paper is written entirely the basis of having a good IV.

What makes a good IV? Obviously, a good IV is one which satisfies the two requirements laid out above.

That is, a good IV is one that is (i) relevant and which (ii) meets the exclusion restriction. Since the latter requirement is the more difficult to satisfy of the two, you should always start with that requirement when looking for an IV.

What Is a Good IV?

Here, it is worth spending time thinking about what it means for an IV to meet the exclusion restriction. Remember from the previous lecture that there are three sources of statistical endogeneity:

1. Reverse causality, or simultaneity,
2. Unobserved heterogeneity, or omitted variables, and
3. Measurement error.

So a good IV is one that will take care of all of those problems... and then some. In other words, you need to argue that your IV takes care of all three of those problems (or that some of those problems are not an issue in your application), and *then* you need to explain that your IV only affects your outcome variable through the variable of interest. In a few words: IV is hard!

What Is a Good IV?

Once you have managed to convince yourself (and, more importantly, your readers or your audience) that your IV meets the exclusion restriction, you need to pray that it is relevant.

With a weak IV, Bound et al. (1995) point out that using IV can be worse than just using plain old OLS: with a weak IV, your IV estimates are biased toward OLS estimates, and your IV estimates may not even be consistent. Moreover, significance tests have the wrong size, and confidence intervals are wrong. In other words, both your point estimates and your standard errors get muddled.

I have fortunately never had to deal with a weak IV, but if I had to deal with one, my advice would be to drop it and use a different IV.

What Is a Good IV?

Also note that *you cannot test for exogeneity*.

Some people think that they can simply run a Hausman test in order to ensure that D is exogenous to y. In order to do so, however, the Hausman compares IV results with OLS results. If the IV is bad, the test is worthless. If the IV is good, the best-case scenario is for the Hausman test to reject the null of exogeneity, in which case the IV is needed, and so the test was pointless to begin with.

The worst-case scenario is a failure to reject the null, which is not a very powerful result, as discussed in the previous lecture. In cases where you fail to reject the null of the Hausman test with a good IV, my advice is to present both the OLS and the IV results, and discuss what each estimator generates.

What Is a Good IV?

Worse, because the OLS and IV/2SLS estimators estimate different things (ATE and LATE, respectively), even with a good IV, the Hausman test compares apples and oranges.

That is, the parameter vectors compared in a Hausman test pitting IV/2SLS results against OLS are different by construction, leading to over-rejecting the null of exogeneity.

One might as well conclude that the Hausman test is useless in this context.

Regressions as Ecosystems

It is not uncommon for people with only a passing knowledge of applied econometrics to ask questions of the form "I am studying the effect of D on y; what's a good instrument?"

Whenever you get asked that question, your answer should be "What's in x?" To see why, consider the fact that a regression is an ecosystem, and that all the pieces matter.

This is especially the case if you don't have an experiment or a quasi experiment, and you have to rely on an instrumental variable (IV) that is nonrandom.

Regressions as Ecosystems

Put differently, an IV lives and dies by the controls it is surrounded with. Indeed, here is something that I bet is taking place almost daily throughout the world in economics seminars:

1. The presenter is interested in the causal relationship flowing from some treatment D to some outcome y.
2. The presenter recognizes that y and D are jointly determined, and is thus using an instrument z to get at it.
3. A clever member of the audience says: “Yes, but have you considered [channel through which z violates the exclusion restriction]?”
4. The presenter says: “You're right in principle. Because I have [specific variable] in my set of controls x, the exclusion restriction is still met.”
5. Clever member of the audience: “Oh, okay. Go on then.”

Regressions as Ecosystems

For example, in Bellemare (2015), I was interested in the causal effect of food prices on the extent of social unrest, si I used natural disasters as an IV for food prices.

A few times in seminars, I was asked: “Yes, but you don’t control for the income of food consumers, and that causes omitted variables bias.”

My response was: “Yes, but I am regressing on the real—not nominal—price of food, which controls for the overall price level and thus, presumably, for wages, which themselves determine most people’s income levels.” Thus, when thinking about causality, one should consider $y = f(D(z, x), x) + \epsilon$ as a whole, and not just $D(z)$ or $y = D(z)$.

Regressions as Ecosystems

As a side note, notice how you do not have a choice of which variables to include in equation 2. That is, you cannot pick and choose which controls should be included in your first-stage regression.

The way IV works is as follows: z serves as an IV for D, and each element of x serves as an IV for itself.

This does not mean the number of regressors in the first and second stage need to be equal; it is not impossible for you to have more variables in z than you have variables in D. Should you be so fortunate, you can run tests of overidentification restrictions. What this means, however, is that you do not get to cherrypick which control variables will be included in your first-stage regression.

Better LATE than Nothing

It is worth thinking about what it is we are estimating with IV. Recall that our grand aim is to estimate average treatment effects (ATEs).

With IV, however, it is rare that we can estimate an ATE. Rather, we have to settle for a local average treatment effect (LATE).

What does LATE tell us about the world? Intuitively, LATE estimates the effect of D on y for the subset of observations (i.e., individuals, households, firms, etc.) which were induced to take up the treatment D in response to a change in z. For the remainder of this section, let's consider the case where both z and D are dichotomous variables.

Better LATE than Nothing

We call those observations for which $D = 1$ in response to $z = 1$ and those observations for which $D = 0$ in response to $z = 0$ compliers; the other observations are called noncompliers, and we can split them up in two groups: (i) never-takers, i.e., people for whom $D = 0$ no matter what z is equal to, and (ii) always-takers, i.e., people for whom $D = 1$ no matter what z is equal to.

It is sometimes possible to know who the compliers and noncompliers are, but not always. As a result, the precise subset of observations for which your IV results hold for can sometimes be nebulous.

Better LATE than Nothing

With that said, there are two key identifying assumptions underlying LATE: (i) conditional independence, and (ii) monotonicity.

The former simply says that the joint distribution of D and y is independent of z, and is a restatement of the exclusion restriction. The latter is more tricky, as it requires that z should push D in the same direction (or no direction) for all observations.

In other words, it is fine for z not to have an effect on D for a subset of observations, but when it does have an effect on D, that effect needs to be the same for all observations induced by the IV to take up the treatment.

Better LATE than Nothing

Lastly, note that with multiple IVs, precisely what the LATE is becomes very complex.

Suppose you have an IV z_1. A first discrepancy between ATE and LATE is introduced because of compliance issues regarding z_1—some observations are induced to take up treatment D by z_1, others not, and the latter mess up one's estimate of the ATE.

But then, suppose we introduce a second instrument z_2. That introduces a whole new compliance issue—some observations were induced to take up treatment D by z_2, others not, and the latter mess up one's estimate of the ATE, too. It can be difficult to think about the potentially overlapping sets of compliers and noncompliers.

RCTs and IV

One case where the IV setup can be extremely useful is when you have an RCT with imperfect compliance—say, because you cannot force people to take up a treatment.

For example, I once ran an RCT in which we randomly provided cotton producer cooperatives in Mali with an index insurance product which paid out in case of low area yields (Elabed et al., 2013).

Because insurance was an unknown financial product for most of those producers, and since coops would have to pay to be insured, it would have been unethical to force them to take up the insurance.

RCTs and IV

So what we did instead was to offer an encouragement design: cooperatives were offered a discount of 25, 50, or 75 percent at random on the insurance.

Because we wanted to know the effect of the insurance treatment D on a number of welfare measures y, we used the random discount as an IV z, which allowed us to estimate the LATE.

In this case, the monotonicity assumption was satisfied because as the size of the discount increase from 25 to 50, and from 50 to 75 percent, it was extremely unlikely that people would become less likely to buy the insurance—at worst, the discount would not have changed their minds. So while it was not possible for us to estimate an ATE, the LATE was still useful.

Exogenous to What?

One of the worst things you can do as an applied microeconomist is to unthinkingly re-use someone else's IV, without making sure that the IV actually works in your application; this is a corollary of the regressions-as-ecosystems discussion above.

One of the IVs that has gotten overused in recent years—to the point where it eventually became a punchline—is rainfall. After all, the (mistaken) reasoning goes, rainfall is exogenous, because there is no way on earth your variable of interest actually causes rainfall, right?

Exogenous to What?

There are two mistakes with that reasoning:

1. “Endogeneity” is about statistical—not theoretical—endogeneity. Both unobserved heterogeneity and measurement error also are causes of statistical endogeneity.
2. Exogenous to what, exactly? That is, an instrumental variable z which you use to identify the causal impact of a treatment variable D on some outcome y will (i) work only if it is exogenous to the outcome y, i.e., if it only affects y through D, and (ii) lives or dies by the controls x it is surrounded by. Sometimes, an IV will only work if you use your controls x wisely to eliminate potential channels through which the exclusion restriction is violated.

And as you would expect, someone came along demonstrating that rainfall is not the magical IV some would have liked it to be (Sarsons, 2015).

Bellemare (2015)

In this paper, I was interested in the effect of food prices on social unrest.

To see whether food prices actually caused social unrest, I assembled a data set including a measure y of social unrest (a count of all news stories in the English-language media reporting instances of social unrest) and a measure p of food prices (the FAO's food price index), both worldwide.

The data I assembled covered the period 1990-2011, for a total of 262 monthly observations.

Bellemare (2015)

Though the data are time series data, I used the standard applied micro toolkit.

Specifically, because food prices can be endogenous to social unrest, I relied on an IV setup to estimate the causal effect of food prices on social unrest.

The IV I relied on was a count of natural disasters worldwide in a given month.

Bellemare (2015)

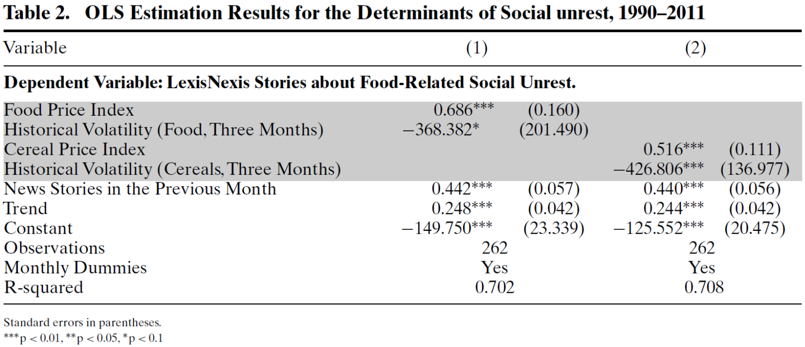


Table 2. OLS Estimation Results for the Determinants of Social unrest, 1990–2011

Variable	(1)		(2)	
Dependent Variable: LexisNexis Stories about Food-Related Social Unrest.				
Food Price Index	0.686***	(0.160)		
Historical Volatility (Food, Three Months)	−368.382*	(201.490)		
Cereal Price Index			0.516***	(0.111)
Historical Volatility (Cereals, Three Months)			−426.806***	(136.977)
News Stories in the Previous Month	0.442***	(0.057)	0.440***	(0.056)
Trend	0.248***	(0.042)	0.244***	(0.042)
Constant	−149.750***	(23.339)	−125.552***	(20.475)
Observations	262		262	
Monthly Dummies	Yes		Yes	
R-squared	0.702		0.708	

Standard errors in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Bellemare (2015)

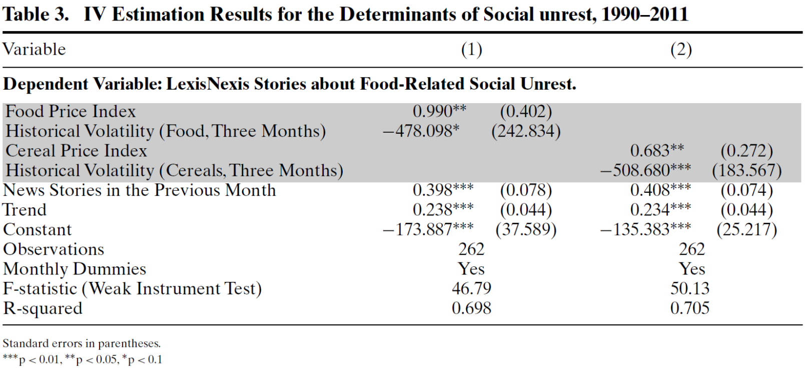
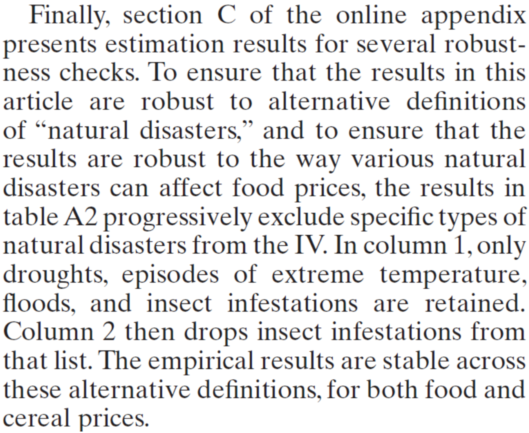


Table 3. IV Estimation Results for the Determinants of Social unrest, 1990–2011

Variable	(1)		(2)	
Dependent Variable: LexisNexis Stories about Food-Related Social Unrest.				
Food Price Index	0.990**	(0.402)		
Historical Volatility (Food, Three Months)	−478.098*	(242.834)		
Cereal Price Index			0.683**	(0.272)
Historical Volatility (Cereals, Three Months)			−508.680***	(183.567)
News Stories in the Previous Month	0.398***	(0.078)	0.408***	(0.074)
Trend	0.238***	(0.044)	0.234***	(0.044)
Constant	−173.887***	(37.589)	−135.383***	(25.217)
Observations	262		262	
Monthly Dummies	Yes		Yes	
F-statistic (Weak Instrument Test)	46.79		50.13	
R-squared	0.698		0.705	

Standard errors in parentheses.
***p < 0.01, **p < 0.05, *p < 0.1

Bellemare (2015)



Finally, section C of the online appendix presents estimation results for several robustness checks. To ensure that the results in this article are robust to alternative definitions of “natural disasters,” and to ensure that the results are robust to the way various natural disasters can affect food prices, the results in table A2 progressively exclude specific types of natural disasters from the IV. In column 1, only droughts, episodes of extreme temperature, floods, and insect infestations are retained. Column 2 then drops insect infestations from that list. The empirical results are stable across these alternative definitions, for both food and cereal prices.

Summary

- You need to defend your IV. Even with an experimental IV (e.g., the encouragement design we used in our study of insurance for cotton producers in Mali), you need to explain why your IV meets the exclusion restriction, and you need to take each source of statistical endogeneity in turn and explain why the IV works against it (or why that source of endogeneity is not an issue in your application). Anything less, and your paper will not only be likely to be rejected—it will deserve to be rejected.

Summary

- You need to show both your first- and second-stage equations. Though the latter is always included, some people omit the former, which might lead some readers to wonder what those people are hiding.

Summary

- You need to show the F-statistic on your IV in the first-stage equation. This is second in importance to your discussion of the validity of your instrument, and omitting it will cause reviewers to recommend a rejection.
- It is always a good idea to show a regression of y on z, i.e., a reduced-form regression of your outcome of interest on the IV. This serves as a first test that the IV is actually significantly correlated with y. If there is no statistically significant relationship between the two, you will need to spend time thinking about and discussing why that is the case. As Angrist and Pischke (2009) note, "if you can't see the causal relation of interest in the reduced form, it's probably not there" (also see Chernozhukov and Hansen, 2008).

Summary

- Here is a trick I picked up from Acemoglu et al. (2001): Because the relationships of interests when doing IV are between (i) y and D, (ii) D and z, and (iii) y and z, it is always useful to show scatterplots of those relationships overlaid with univariate regressions thereof. Though economists tend to be very critical, it is almost as if just showing those graphs can convince your readers that the relationship you are after is truly there in the data, provided those graphs actually show this convincingly. In other words, there is often something to a result that can just be eyeballed which makes anything else—all the discussion surrounding—mere icing on the cake.

Summary

- The line between IV and control can sometimes be difficult to find. In Bellemare (2012), I successfully made the case that respondent WTP to participate in contract farming should be an IV for participation in contract farming, making the theoretical argument that preferences determined the decision to participate. After a colleague pointed out that it might work better as a control in a selection-on-observables design, in Bellemare and Novak (2016) and Bellemare et al. (2018), we do just that.

Summary

- You should always compare and contrast your IV results with a naïve OLS specification. With a good (i.e., credible and relevant) IV, if you find that there is little difference between the LATE and the ATE, it is quite possible that endogeneity is not a huge issue in your application. Yet you should always argue from a position of methodological skepticism, and use utmost caution in making the case that something is exogenous. It might well be endogenous, but not detectably so.

Summary

- In Bellemare (2015), when I used natural disasters as an IV for food prices in a regression of food riots on food prices, I had the luxury of being able to include or exclude different categories of natural disasters from my IV. In order to assess the robustness of my core findings, I estimated specifications with different types of natural disasters as IVs.

Summary

- When all is said and done, if you have to use IV, stick to linear regression. This means that if you have a binary dependent variable, instead of estimating a fancy model like IV probit, you should just stick to 2SLS. This is because linear models are clean and clear, and they cannot lead to identification via functional form, as is often the case with some ML-based procedures. Again, nowadays, what matters is credibility, and properly modeling the DGP is secondary.

Summary

- You might be tempted to use lagged values of endogenous variables as IVs. In Bellemare et al. (2017), we show that lagged control variables only lead to identification under a set of assumptions that is extremely stringent; this is also true for lagged variables as IVs.
- When writing an IV-based paper, it is best to include a section where you entertain the various ways in which your exclusion restriction could be violated, and either discuss how either do not apply (say, because you have included the right controls) or how show additional results where you include the relevant "backdoor" channels, as per Pearl's terminology, and show that they do not weaken your identification.

Summary

- Think carefully about who the compliers and noncompliers are in your application, and discuss who the LATE applies to.
- Think carefully about whether your IV has a monotonic effect on the treatment variable. If it does not, be honest.

Summary

- Avoid estimating regressions where you have two endogenous variables. It is hard enough to identify one causal relationship, try to keep the other relationship of interest for a separate paper. Your troubles will grow exponentially the more endogenous variables you include.

Summary

- Recall that $\widehat{D}_i$ is a generator regressor. As such, if you were to do 2SLS by hand (and there really is not much of a reason to do so), you would need to correct your standard errors. Luckily, bootstrapping them works in this case and is easily implemented with most statistical packages.
- Always remember that 2SLS is unbiased only asymptotically. In small samples, 2SLS is consistent but biased. This is particularly true with weak IVs, and the size of the bias increases with the number of IVs.