Causal Inference with Observational Data

2. Causality

Marc F. Bellemare

May 2018

Causality

I begin this class with a discussion of causality because for all intents and purposes, (getting as close as possible to) identifying causal relationships is what the vast majority of applied microeconomists spend their time working on.

Even the staunchest of structural econometricians, whose time is usually not spent thinking about clever identification strategies, is usually interested in whether the exogenous variables in her models cause the endogenous variables.

(It is important not to confuse the theoretical and empirical definitions of exogeneity and endogeneity. Many disagreements and misconceptions stem from those homonyms. More in this in a minute.)

Causality

And for what it's worth, it's not just economists who care about identifying causal relationships—since I got my PhD in 2006, I have seen a methodological convergence take place in the social science.

My one purely econometric contribution—a paper in which my coauthors and I show that lagging explanatory variables will generally not exogenize them—was published in the *Journal of Politics*, and for my money, the methodological pieces published in the top political science journals are often a lot more useful to my work than those published in the top economics journals.

Likewise, social scientists in sociology, criminology, etc. are doing quantitative work with the goal of identifying causal relationships (Manzi, 2010).

Causality

But before delving into causality, I should discuss two things about myself which explain where my point of view comes from.

First, when I was doing my bachelor's degree at the Université de Montréal in the late 1990s, the only social science students who took any serious statistics were economics majors. Students who majored in political science, sociology, or anthropology never took any math or stats classes.

In other words, social sciences at the Université de Montréal were done the old-fashioned French way.

Causality

Second, toward the end of my Master's degree, I dated someone who was writing her doctoral dissertation on the relationship between a large multinational corporation and its employees—a large corporation *which she used to work for*.

Worse, her structured interviews relied on a convenience sample *of the friends she had made while she worked for that corporation*.

So unlike the current generation of graduate students, my reference point for what constitutes rigorous empirical work in the social sciences was set extremely low to begin with.

Causality

Suppose we have the following theoretical relationship:

$$y = f(x). \tag{1}$$

This relationship is deterministic—if we know x and $f(\cdot)$, we know y.

Beyond being deterministic, the relationship above might be causal: By saying that y is a function of x, we are implying that x causes y.

Causality

In other words, because convention dictates that y should be on the left-hand side (LHS) and x on the right-hand side (RHS) of equation 1, we suspect that causality flows from x to y in the same equation. Nothing, however, prevents us from writing the same equation as $x = f^{-1}(x)$.

Suppose we have data on y and x for a sample of observations $i = 1, ..., N$. Those variables need not be the same as in the equation above. Linearly projecting y on x yields

$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{2}$$

where the error term ϵ_i is added because the relationship is now stochastic rather than deterministic.

Causality

Now suppose we estimate equation 2, ignoring for the moment how we do that (i.e., ordinary least squares, maximum likelihood, or method of moments).

Is the coefficient estimate $\widehat{\beta}$ causally identified? (Let's ignore for the time being the imprecisions of language surrounding the terms "causal" and "identified.")

The answer is "Maybe, but probably not."

Causality

For starters, we know that $\widehat{\beta}$ is identified (or unbiased) if and only if $Cov(x, \epsilon) = 0$, in which case $E(\widehat{\beta}) = \beta$.

But even then, suppose you know for a fact that $Cov(x, \epsilon) = 0$ (say, because you randomly assigned x). A true skeptic might not buy your story, simply because who knows if you're not in the one in 10, 20, or 100 cases (depending on your level of confidence) where you reject the true null hypothesis that $\beta = 0$?

For a true skeptic, accepting any statement as causal requires no less of a leap of faith than that necessary to believe in the existence of a Great Architect of the Universe.

Causality

Luckily, in economics and other social sciences, the claims we are trying to find causal evidence for are far removed from theological claims.

Specifically, if you can convincingly argue on the basis of your research design that $Cov(x, \epsilon) = 0$ (arguably a big "if"), then provided that you did not make any mistake in estimation, then your estimated relationship is causal beyond any reasonable doubt.

Causality

In other words, the name of the game in the applied micro version of econometrics is to estimate a version of

$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{3}$$

where we take as much "bad" stuff out of ϵ as possible, with "bad" stuff defined as "stuff that is correlated with x." This is what the title of this course—causal inference with observational data—refers to.

Corollary: Any causal claim is based on a selection-on-observables argument.

Bellemare and Novak (2017)

Selection on observables example.

We were interested in the impact of participation in contract farming (D) on the duration of the hungry season y for household i holding a number of potential confounders x constant. So we are interested in the coefficient γ in the equation

$$y_i = \alpha + \beta x_i + \gamma D_i + \epsilon_i. \tag{4}$$

The issue is obviously that participation in contract farming is not randomly assigned to the households in the data—households choose (not) to participate on the basis of things which we typically do not observe (e.g., ambiguity aversion, discount rates, entrepreneurial ability, managerial ability, risk aversion, technical ability, etc.)

Bellemare and Novak (2017)

Luckily, the survey questionnaire included a series of questions aimed at eliciting (all) respondents' WTP to participate in a hypothetical contract farming arrangement which would increase their income by 10%. Considering the following Roy model (Smith and Sweetman, 2016), household i participates iff

$$y_{1i} - c_i \geq y_{0i}, \tag{5}$$

where c_i is i's cost of participation. For each household, we know y_{0i}, and I also know $y_{1i} = y_{0i} + 0.1y_{0i}$.

Bellemare and Novak (2017)

By exogenously varying c_i in the in-survey experiment and observing people's yes or no answers to the hypothetical question, we can obtain for each respondent a measure of his WTP for participation in the hypothetical arrangement (let's ignore how I do so in the interest of brevity), which proxies for his marginal utility MU of participating in contract farming. This means I can estimate

$$y_i = \alpha + \beta x_i + \gamma D_i + \delta MU_i + \epsilon_i. \quad (6)$$

Since a respondent's marginal utility will be moved around by typically unobservable factors (e.g., ambiguity aversion, discount rates, entrepreneurial ability, managerial ability, risk aversion, technical ability, etc.) then including a proxy for it should pull out of the error term those factors which account for selection.

Bellemare and Novak (2017)

Table 2a. Estimation Results for OLS, Cox Proportional Hazard, and Survival-time Regressions Omitting WTP Variables

Variables Dependent Variable: Duration of Hungry Season	OLS	Cox	Survival Time
Contract farming participant	−0.294** (0.142)	0.150** (0.062)	0.171** (0.070)

Bellemare and Novak (2017)

Table 2b. Estimation Results for OLS, Cox Proportional Hazard, and Survival-time Regressions Including WTP Variables

Variables Dependent Variable: Duration of Hungry Season	OLS	Cox	Survival Time
Contract farming participant	−0.277* (0.145)	0.166*** (0.063)	0.188*** (0.071)

Causality

What makes applied econometrics more art than science—more rhetoric than dialectic—is the fact that one cannot test for causality.

Rather, one has to argue that one's research design yields a causal relationship. How easy this is depends in large part on your research design.

The reason I teach a course such as this one is that most graduate programs are better at teaching you how to run tests than they are at teaching you rhetorical skills!

Causality

Before anything, it should go without saying that identifying a causal relationship flowing from x to y does not mean that y is only caused by x.

In Bellemare (2015), I showed beyond any reasonable doubt that rising food prices levels cause food riots. At a policy conference in Washington, DC, I was taken to task by another participant—a physicist—for talking about causality... because food riots have more causes than just food prices.

Well, yeah. That isn't the point of quantitative social science!

Causality

The latter claim is hard to dispute—when food prices go up, the fact that we are more likely to see food riots in Lagos than in New York City is clearly proof of that—but one has to be careful not to interpret the statement "x causes y" as equivalent to the statement "x is the only cause of y."

The former can be identified with a good research design; the latter is akin to the Unmoved Mover of Aristotle's *Metaphysics*.

Statistical Endogeneity

What makes $Cov(x, \epsilon) \neq 0$? That is, what are the sources of statistical endogeneity? Broadly speaking, it is useful to break things down into three such sources.

The first is *reverse causality,* or *simultaneity*. This arises when x causes y but y also causes x. If your observations cover a long enough time span, this is likely to happen. Alternatively, one might say that the expectation of y might cause individuals (or firms, or households, etc.) to adjust x consequently.

In a regression of wage on education, for example, it is almost certain that individuals' expectations regarding their future wage has driven how much education they have gotten.

Statistical Endogeneity

The second source of statistical endogeneity is *unobserved heterogeneity*, or *omitted variables*.

In most applications in applied microeconomics, this is the main source of statistical endogeneity. Individuals' preferences, levels of ability, etc. are typically not observed by the econometrician, and they are likely to be correlated with what the econometrician can observe, in particular the variable of interest.

Statistical Endogeneity

The third source of statistical endogeneity is *measurement error*.

This arises when one of your variables—in particular, your variable of interest—is systematically misreported or mismeasured, and the degree of misreporting or mismeasurement is correlated with what you can observe.

Note: This is distinct from classical measurement error, where something is measured with error at random.

Statistical Endogeneity

In all three cases, there is something in the error term ϵ in equation 2 which is correlated with x, which means that $Cov(x, \epsilon) \neq 0$ and $E(\widehat{\beta}) \neq \beta$.

Why do I talk of *statistical* endogeneity?

Because there is a vast difference between theoretical and statistical endogeneity. Theoretical endogeneity (exogeneity) refers to the case where the value of a variable is determined (taken as given) within a specific optimization problem. Statistical endogeneity (exogeneity) refers to cases where $Cov(x, \epsilon) \neq 0$ ($Cov(x, \epsilon) = 0$).

Statistical Endogeneity

The two notions have little in common with each other.

Unfortunately, the fact that we use the same term is confusing, and some economists—in particular, those who were trained before Credibility Revolution (Angrist and Pischke, 2010) and failed to catch up on empirical methods—mistakenly believe the two to be identical.

This leads to some people thinking of reverse causality to be the definition of statistical endogeneity. It is not; it is only one cause of statistical endogeneity.

Statistical Endogeneity

The foregoing suggests a systematic way to think through and discuss identification issues when writing applied papers.

In my own applied work, I almost always include a point-by-point discussion of whether (i) reverse causality or simultaneity, (ii) unobserved heterogeneity or omitted variables, and (iii) measurement error are a source of bias in the application at hand, and of how I deal with those sources of statistical endogeneity that are there.

I have come to see such a discussion as second only to an article's introduction in terms of importance, and I believe most young researchers would benefit from including such a discussion when using observational data.

Methodological Skepticism

David Hume (1711-1776) was one of the many philosophers of science who carefully thought and wrote about causality. According to Lorkowski (2016),

> *[I]f the denial of a causal statement is still conceivable, then its truth must be a matter of fact, and must therefore be in some way dependent upon experience. Though for Hume, this is true by definition for all matters of fact, he also appeals to our own experience to convey the point. Hume challenges us to consider any one event and meditate on it; for instance, a billiard ball striking another. He holds that no matter how clever we are, the only way we can infer if and how the second billiard ball will move is via past experience. There is nothing in the cause that will ever imply the effect in an experiential vacuum.*

Methodological Skepticism

Extrapolating from the last two sentences to economics, for Hume, a good theoretical model is of no help in identifying causal relationships.

Worse, a theoretical model is completely useless without at least *some* data (this could be something as simple as stylized facts) to test it.

Methodological Skepticism

With the Credibility Revolution, applied microeconomists have adopted a position of methodological skepticism.

That is, before the average applied microeconomist can take a given relationship as causal, she has to be convinced of it.

The default position is to assume that any given correlation is just that—and not a causal relationship.

Methodological Skepticism

When a researcher claims that a given relationship is causal, the onus is on her to prove it beyond any reasonable doubt.

It is very much in this sense that much of applied microeconomics is a craft, and that much of our work is rhetorical: In the absence of an experiment or quasi experiment, it is difficult to claim that a given relationship is causal.

So when someone is skeptical of another's identification strategy at a seminar, this is (usually) not because the former person is being obnoxious. Rather, it is because that person is merely exhibiting the kind of methodological skepticism which (for better or for worse) is equated with critical thinking nowadays in our profession.

Methodological Skepticism

One problem is that you cannot test for endogeneity.

You can test for exogeneity—that is, you can run a test that assumes there is no statistical endogeneity, as in the Durbin-Wu-Hausman test—but a failure to reject the null in such cases is not convincing: With 90, 95, or 99 percent of the probability mass resting on the null, depending on your chosen level of confidence, you would expect to fail to reject the null in most cases.

Thus, a rejection of the null in this case is much more convincing than a failure to reject the null. The problem is that most people who run Durbin-Wu-Hausman tests are usually interested in "proving" there is no endogeneity. But it is difficult to prove a negative—you could spend a lifetime trying to prove that unicorns do not exist.

Pearl's Contribution

A lot of ground has been covered since the days of David Hume when it comes to the study of causality. The leading researcher on causality nowadays is Judea Pearl, a computer scientist at UCLA. One of Pearl's insights is that we simply do not have the notation to talk about causality.

Let us take equation 2 again. What we are interested in is in estimating $P(y|x)$, i.e., the probability that y will take a given value given that we know the value of x, which is such that

$$P(y|x) = \frac{P(y,x)}{P(x)}. \tag{7}$$

Pearl's Contribution

The problem is that equation 7 tells us nothing about whether the relationship between y and x is causal! We could also write

$$P(y, x) = P(x|y)P(y), \tag{8}$$

which also tells us nothing about causality. This is equivalent to saying that equation 2 could easily be rewritten as

$$x_i = \pi + \phi y_i + \nu_i, \tag{9}$$

where $\pi = -\alpha/\beta$, $\phi = 1/\beta$, and $\nu = -\epsilon/\beta$. In other words, the same equation can be written in two ways, without there being any indication as to the direction of causality.

Pearl's Contribution

Pearl (2009) suggests that we need a new notation, $do(x)$, which indicates that we "do something" to x. That is,

$$P(y|do(x)), \tag{10}$$

where $do(x)$ indicates that the econometrician controls x in some way (for instance, via an experiment).

Only then can we truly talk of causality.

Pearl's Contribution

Economists have been thinking about causality for a while. In two articles published a half-century ago, Herman Wold discussed the notion of causality in econometrics (Wold, 1954) as well as causal inference in observational data (Wold, 1956).

The study of causality has been neglected in economics until the mid-1980s, if not the early 2000s. Even then, Kennedy (2008) only discussed causality briefly in the context of Granger causality—and then again, to warn the reader that Granger causality is *not* causality because the sales of holiday greeting cards have been found to Granger-cause the holidays.

Pearl's Contribution

Pearl also brought to the study of causality the use of directed acyclic graphs (DAG).

Strictly speaking, a DAG is a finite, directed graph with no directed cycles.

In econometrics, DAGs are used to graph the that some variables have on other variables. As such, DAGs are useful in that they are visual representations of the inference problem at hand, and they can help us determine visually whether an estimated relationship is causally identified or not.

Pearl's Contribution

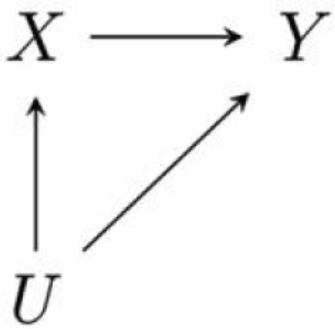

Figure 1. The identification problem. This is a representation of a causal relationship from X to Y where identification is compromised by unobservables U.

Figure: Source: Bellemare, Masaki, and Pepinsky (2017).

Regression vs. Matching

I usually teach applied economists, who are familiar with the regression approach, so that's what I focus on in this course.

But with a research design that allows assuming conditional independence, the matching approach is also valid. One distinct advantage of matching methods is that they sometimes allow estimating types of treatment effects which are otherwise impossible to estimate. In Bellemare and Novak (2017), for example, we rely primarily on a regression approach to estimate the average treatment effect (ATE) of interest, and we then rely on a matching approach (i) to assess the robustness of our regression results, and (ii) to estimate both the average treatment effect on the treated (ATT) and the average treatment effect on the untreated (ATU).

Regression vs. Matching

In terms of notation, assuming a binary treatment variable D, let

$$ATE = E(y_i|D_i = 1) - E(y_i|D_i = 1), \tag{11}$$

with the ATT and ATU defined as

$$ATT = E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 1), \tag{12}$$

where y_{1i} and y_{0i} respectively denote the value of the outcome variable for observation i in cases where i is treated and untreated, and

$$ATU = E(y_{1i}|D_i = 0) - E(y_{0i}|D_i = 0). \tag{13}$$

Bellemare and Novak (2017)

Table 6. Outcome Variable: Duration of Hungry Season

Sample	1 Neighbor Caliper 0.01	3 Neighbors Caliper 0.01	3 Neighbors Caliper 0.001
Unmatched sample	−0.400***	−0.400***	−0.400***
	(0.123)	(0.123)	(0.123)
Average treatment effect on the treated	−0.194	−0.305	−0.295
	(0.234)	(0.223)	(0.255)
Average treatment effect on the untreated	−0.062	−0.204	−0.249
	(0.225)	(0.207)	(0.269)
Average treatment effect	−0.127	−0.252	−0.272
	(0.204)	(0.196)	(0.241)

Note: Standard errors appear in parentheses; standard errors calculated using three neighbors to calculate conditional variance as in Abadie and Imbens (2006). Asterisks denote the following: ***= $p < 0.01$, **= $p < 0.05$, and *= $p < 0.1$.

Regression vs. Matching

Intuitively then, the ATT and ATU measure the causal effect of changing $D = 0$ to $D = 1$ for those who were treated and causal effect the same change would have on those who were not treated.

The issue with matching is that oftentimes, researchers who lack a credible research design will substitute matching on observables for that credible research design and claim that it allows making more credible statements than a regression approach would.

But matching on observables does not account for unobservables, which is usually what plagues economic applications.

What to Tackle, and In Which Order

Frances Woolley wrote:

> *[I]t is rare that I will have someone come to my office hours and ask "Have I chosen my sample appropriately?" Instead, year after year, students are obsessed about learning how to use probit or logit models, as if their computer would explode, or the god of econometrics would smite them down, if they were to try to explain a 0-1 dependent variable by running an ordinary least squares regression. I try to explain: "Look, it doesn't matter. It doesn't make much difference to your results. It's hard to come up with an intuitive interpretation of what logit and probit coefficients mean, and it's a hassle to calculate the marginal effects. You can run logit or probit if you want, but run a linear probability model as well, so I can tell whether or not anything weird is going on with the regression." But they just don't believe me.*

What to Tackle, and In Which Order

Indeed, nothing screams "grad student" louder than an obsession with fancy estimators—usually of the maximum likelihood variety (i.e., probit, logit, tobit, etc.), sometimes of the semiparametric variety—instead of with whether one has reasonably identified one's parameter of interest (via a research design that relies on a plausibly exogenous source of variation), or with whether one's findings have some reasonable claim at being externally valid via the use of a representative sample.

What to Tackle, and In Which Order

There is an ontological order of importance to things in applied work, which unfortunately goes unspoken in most econometrics classes. That order is roughly as follows:

1. *Internal validity*. Is your parameter of interest credibly identified?
2. *Precision*. Are your standard errors right?
3. *External validity*. Are your findings applicable to observations outside of your sample?
4. *Data-generating process*. Did you properly model the DGP?

What to Tackle, and In Which Order

Getting standard errors right is important. But it is not more important than internal validity. At least not these days.

Likewise, it is important to account for the fact that a dependent variable is ordered and categorical, but with 150 observations, one is better off relying on a good research design and using a linear regression than a likelihood-based procedure (which is only asymptotically consistent; $n = 150$ does not count as asymptotic).

Conversely, having "big data" in the form of millions or billions of observations will not make your work more likely to be published in good journals in the absence of solid identification.

Summary

- Unless you are dealing with experimental data, where causality is practically given, start from a position of methodological skepticism.
- Think carefully about what leads to $Cov(x, \epsilon) \neq 0$ in your application. Your paper should have an Empirical Framework section. In that section, which should be split in at least two sub-sections—Estimation Strategy and Identification Strategy—systematically list the three causes of statistical endogeneity in your Identification Strategy section and explain how your research design allows ruling them out as concerns.

Summary

- If your research design does not allow ruling one of those sources of statistical endogeneity out, be honest about it, and try to explain how that source biases your results. Drawing a DAG might help. In some cases, you might be able to analytically derive the sign and magnitude of the remaining bias. Any attenuation bias is good for your story when you reject the null, since it implies that what you have estimated is an estimate of a lower bound on the true effect.
- Another thing which works well is to imagine what the perfect data set to answer your question would look like, explain how the data you use in your paper differs from that ideal, and then explain how given available data and methods, you are as close as possible to the ideal data set.

Summary

- Whatever you do, unless you have experimental or quasi experimental data, or data from a randomized control trial, do not use causal language. Instead of talking about how x causes y, talk about how your results *suggest* that x causes y, about how there is an *association* between the two. Papers get rejected when their authors use causal language where it is not warranted.
- Focus on internal validity, i.e., on identification, first and foremost. Your dependent variable might be a count variable, but as long as you have not done a good job of identifying whether your variable of interest causes it, estimating a Poisson or negative binomial regression remains secondary. At best, you can estimate those fancier regressions as robustness checks.