The Credibility Revolution in Economics: Agricultural and Applied Economists, Take Note

Marc F. Bellemare

Sanford School of Public Policy
Duke University

March 16, 2012

The Credibility Revolution in Economics

Over the past 20 years or so, a "credibility revolution" has taken place in applied microeconomics (Angrist and Pischke, 2010).

The credibility revolution is so named because it has led to applied microeconomists adopting considerably more stringent standards of statistical identification.

More specifically, applied microeconomists now strive to identify *causal relationships* rather than mere correlations.

The Credibility Revolution in Economics

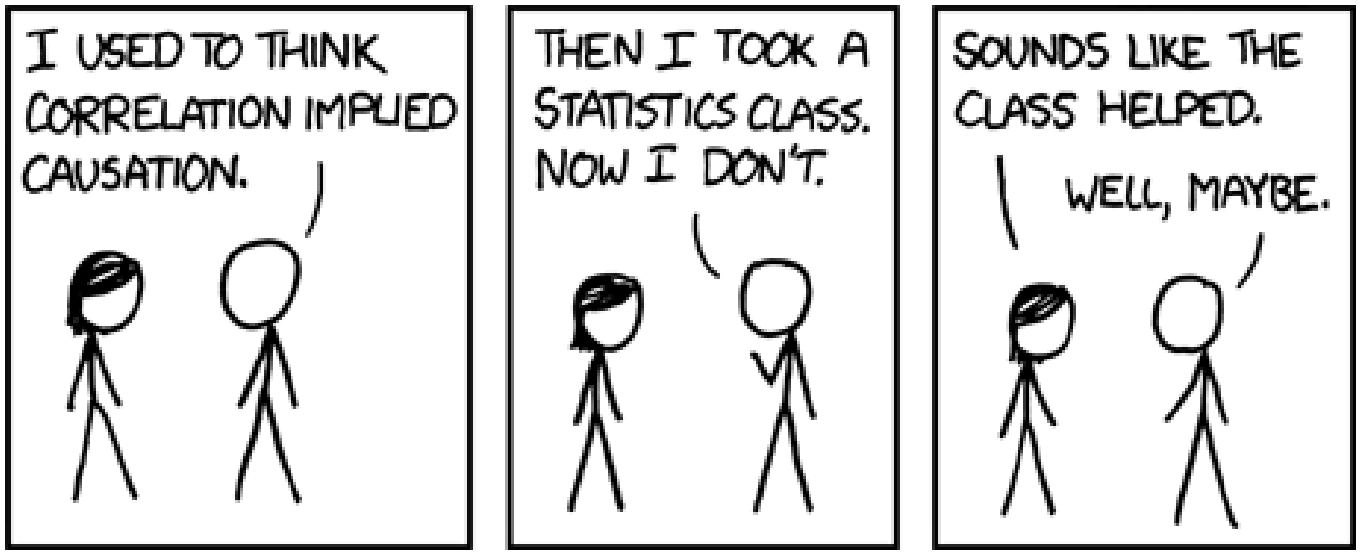

Figure: Correlation $\neq$ Causation (Source: xkcd).

But First, a Confession

Had I given this keynote address three years ago, I would have sung a very different tune.

During the 2009 edition of this conference, I remember disagreeing with a colleague about the importance of identification and diminishing the need for clean identification.

I was wrong. There's a big difference between using the best available research design and focusing only on the questions for which we can have credible causal identification. More on this in the conclusion.

A Necessary Condition for Policy Relevance?

Q: "Any pet peeves with submissions or with referees that would be good for people to avoid?"

A: "Unfortunately yes. Our main two criteria in selecting papers for publication are rigorous identification and policy relevance. The two go together as we cannot have credible policy recommendations without strong causal inference."

– *World Bank Economic Review* editors Alain de Janvry and Élisabeth Sadoulet, in an interview with the Development Impact blog.

What Do We Mean by Identification?

Suppose we have the relationship

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

where y is the outcome of interest, D is a treatment variable (i.e., the variable of interest), x is a vector of controls, and ϵ is an error term with mean zero.

We want to know whether an increase in D *causes* y to change and, if so, by how much.

Under the right circumstances, one could estimate equation 1 and obtain $\frac{\partial y}{\partial D} = \widehat{\gamma}$, where $\widehat{\gamma}$ is causal.

Defining the Problem

Unfortunately, the right circumstances are extremely rare, and the mere act of writing equation 1 as it is written above does not make the estimate of γ causal.

Instead of equation 1, nothing prevents one from writing

$$D_i = \frac{y_i}{\gamma} - \frac{\alpha}{\gamma} - \frac{\beta}{\gamma} x_i - \frac{\epsilon_i}{\gamma}, \tag{2}$$

and theory is often of little to no help in determining whether a relationship is causal.

Defining the Problem

In other words, even though theory is helpful in determining *which research questions to answer*, good empirical work must stand on its own two legs.

That is, the fact that some variable is exogenous in the theoretical model tells us absolutely nothing about whether it is exogenous in the empirical specification of interest.

Worse, in naturally occurring (i.e., observational) data, it is usually the case that everything is endogenous to everything else.

Causal Relationships

What do we need for our parameter estimate $\widehat{\gamma}$ to estimate a causal relationship?

We know from first principles that $E(\widehat{\gamma}) = \gamma$ if and only if $D \perp \epsilon$.

That is, to obtain an unbiased coefficient estimate, it has to be the case that our variable of interest is not just uncorrelated with the error term, it needs to be independent from it.

Sources of Endogeneity

There are three reasons why $D \perp \epsilon$ may not hold, i.e., why we may have a *statistical* endogeneity problem:

1. We may have an omitted variables problem, or unobserved heterogeneity;
2. We may have a measurement error problem;
3. We may have a or reverse causality problem, or simultaneity.

The problem is that too many in the profession still think of endogeneity as stemming only from this last source.

This may be because we ascribe more than one meaning (i.e., theoretical *and* statistical) to the term "endogeneity."

Why Is This a Problem?

First and foremost, this matters because as agricultural and *applied* economists, our work often serves to inform policy. But policy making is costly, especially when it is based on mistaken claims.

Second, this matters because the credibility revolution is being taken seriously by the other social sciences (e.g., criminology, political science, social psychology, sociology, and so on).

If we are not careful, agricultural and applied economics runs the risk of becoming a fringe discipline within the social sciences.

Why Is This a Problem?

This is made all the more important by the current context of budget austerity, wherein

1. Knowing where we can best spend the little money available for policy is of crucial importance, and
2. Some agricultural or applied economics departments have been forced to close.

Herman Wold's Near-Forgotten Contribution

Economists have been thinking about causality for a long time.

In two articles published more than 50 years ago, Herman Wold discussed the notion of causality in econometrics (*Econometrica*, 1954) and causal inference from observational data (*Journal of the Royal Statistical Society*, 1956).

Yet those two articles were not terribly influential – the best-cited of the two has garnered only 232 Google Scholar citations in almost 50 years – and it's safe to say they are not part of the core applied economics curriculum.

The RAND Health Experiment

Between 1971 and 1982, the RAND Corporation ran a large-scale randomized controlled trial (RCT) that randomly assigned 5,809 individuals to different treatments (i.e., health insurance plans) in order to assess subsequent behaviors and outcomes.

Forty years on, the RAND Health Experiment is still hailed as a gold standard (Levy and Meltzer, 2008) because randomization made D (i.e., an individual's given health insurance plan) fully exogenous to y (i.e., an individual's subsequent behavior or health outcome).

In other words, randomization guaranteed that $D \perp \epsilon$.

Lalonde (1986)

Lalonde (1986) made what was probably the first credible attempt at making a causal statement.

In his famous study, Lalonde studied the impact of training programs D on earnings y. The innovation was that individuals were randomly assigned to treatment (i.e., training) and control groups (i.e., no training) groups.

He then tried to compare the results of that RCT (experimental) with those that one would obtain from an observational study, finding that the econometric results did not replicate the experimental results.

Angrist (1990)

But one does not necessarily need an experiment to credibly identify causal relationships.

Angrist (1990) wanted to know the impact of education D on earnings y. The problem, of course, is that education is endogenous to earnings. People choose to acquire education on the basis of unobservables, and future earnings condition how much education one acquires.

So Angrist used individuals' Vietnam draft lottery number as an instrumental variable (IV) for education. Lower lottery numbers mean a greater likelihood of getting drafted, which translates into a greater likelihood of being educated, given the provisions of the GI Bill.

The Rise of Randomized Controlled Trials

The last decade has seen the rise and increased use of RCTs to evaluate various policy ideas (Duflo et al., 2007).

The idea behind a lot of Duflo's work is simple. In order to study whether something (i.e., the treatment) works, you simply adopt Lalonde (1986) identification strategy of randomly assigning subjects between the treatment and control groups. After a specified period of time, you compare outcomes between the treatment and the control group.

For her contribution, Duflo who won the John Bates Clark medal in 2011.

The Development of Field Experiments

The last decade has also seen the rise and increased use of field experiments to test economic theory (Levitt and List, 2009).

The idea behind a lot of List's work is also simple. In order to test some part of economic theory, you run experiments in the field like you would in a lab.

Those experiments aim at eliciting some "deep" parameter (e.g., risk aversion, discount rate, trust, etc.) and then test whether that parameter has the effect posited by the theory on some outcome.

Why Did Economists Forget Causality?

Following Pearl (2009), the idea of identifying causal relationships has likely been forgotten because we simply do not have the mathematical language necessary to discuss it.

Using the notation in equation 1 and ignoring the vector of control variables x, we know that

$$P(y|D) = \frac{P(y, D)}{P(D)}, \tag{3}$$

but this does not tell us anything about whether the relationship between y and D is causal – we can also write that $P(y, D) = P(D|y)P(y)$, which tells us nothing about causality.

Why Did Economists Forget Causality?

Pearl (2009) suggests the new notation $do(D)$ to indicate the fact that we do something to D, such that:

$$P(y|do(D), x), \tag{4}$$

where $do(D)$ indicates that the researcher controls D in some way and that the relationship between y and D is causal. More on what "doing something to D" could mean in a minute.

But first, what are some reactions in agricultural and applied economics to the new standards of identification?

Propensity Score Matching

In Bellemare (forthcoming), I look at the impact of participation in contract farming D on welfare y using cross-sectional data. In an attempt to make a causal statement, I use an experimentally derived measure of respondent willingness to pay to participate in contract farming as an IV.

This is not perfect, but the IV controls relatively well for unobserved heterogeneity, since it captures changes in respondent marginal utility.

A referee was adamant that propensity score matching (PSM) can be used to make a causal statement. I checked, and PSM yields an impact that is six times as high as that estimated via IV.

Propensity Score Matching

In Barrett et al. (2012), a referee asked us I to discuss the possible means of identifying causal relationships one could use in studying the problem we write about.

Taking a first stab at the revisions, I included a discussion of RCTs, IVs, panel data, and regression discontinuity design. A coauthor was adamant that we should discuss PSM methods. But since when can we use x as an IV for D?

Chris Blattman talks of "the cardinal sin of matching": "Matching is not a solution to your endogeneity problem; it is a weighting scheme. Saying matching will reduce endogeneity bias is like saying that the best way to get thin is to weigh yourself in kilos."

IVs from the 1990s

As a referee, I often see questionable instruments used with little to no justification for why they are valid instruments.

What I have in mind is something like using a variable that obviously affects the dependent variable through channels other than the dependent variable. As a hypothetical example, think of using the number of individuals in a household as an IV for that household's participation in contract farming.

A colleague once referred to such identification strategies as using "IVs from the 1990s": there was a time when such IVs were acceptable; they no longer are.

Back to the Future

Figure: "Roads? Where we're going, we can't instrument with roads."

Back to the Future

So what constitutes a valid IV? We are all familiar with the two requirements that an IV be (i) strong, i.e., correlated with D in the first-stage regression, and (ii) plausibly exogenous to y.

But I would like to emphasize an important, tacit part of the second requirement: A valid IV must *only* affect y through D.

This is too often given short shrift, possibly because too many people still think of (statistical) endogeneity as being caused only by reverse causality.

Mechanical Identification via Fixed Effects

Although they help, fixed effects rarely ever solve an endogeneity problem. This is especially true in long panels. There is a tradeoff between T and how much unobserved heterogeneity is purged from the error term.

The fact that fixed effects are not a cure-all is the reason why so few people buy the results of cross-country regressions nowadays – the identification is mechanical.

(And let's not get started on random effects and the weakness of Hausman tests...)

Faux Endogeneity

Lastly, I was told the following anecdote by someone who runs RCTs for a living.

Him and his coauthors submitted a paper to a top journal in agricultural and applied economics in which they randomized D so as to estimate the causal impact of D on y. One of the referees told them: "My main reason for recommending rejection is because D is endogenous to y."

Fortunately, the editor in charge of the paper understood the issue and told them to dismiss that comment.

Meanwhile in Macro

When Sargent and Sims won the 2011 Nobel prize for economics, much was made of how their work helps with the identification of causal relationships.

Not so, says Kevin Grier: "Causality requires identification. Vector autoregressions do not provide any automatic or free identification. To do policy analysis with a VAR (as opposed to agnostic forecasting) one has to make the same type of untestable identifying assumptions here as one does in the older, explicitly simultaneous equation, Cowles commission approach. (...) VARS are not a free lunch."

The Silver Lining

The good news is that an increasing number of agricultural and applied economists realize the importance of identification.

To take two recent examples, Pouliot (2011) and Sneeringer and Key (2011), both published in the *American Journal of Agricultural Economics*, go to great lengths to discuss the identification of their results without running RCTs.

Likewise, Barrett and Carter (2010) offer a thorough discussion of the shortcomings of RCTs but acknowledge the usefulness of better identification.

This list is not exhaustive.

Quo Vadis?

Figure: *Domine, Quo Vadis?*, Annibale Carracci, c. 1602.

Quo Vadis?

So what can we do to remedy the situation? Obviously, we should want our students to be equipped with the requisite critical thinking skills.

But charity begins at home, so the best we can do is to read, in approximate order of importance

1. Angrist and Pischke's (2010) *Journal of Economic Perspectives* article,
2. Angrist and Pischke's (2008) *Mostly Harmless Econometrics*,
3. Shadish et al.'s (2001) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, and
4. Pearl's (2009) *Causality*.

Quo Vadis?

But if reading books and articles is one thing, assimilating their content is another. How can the knowledge gleaned from the sources mentioned above be used in one's own research?

When writing papers, it always helps to start the data section with what the ideal data set would look like in order to answer the question "How does y change when D changes?" or, more broadly, "Does D cause y?"

And usually that first-best data set involves random assignment of observations to different values of D.

Quo Vadis?

But randomization is not always an option. So carefully explaining why one's data set is second-best (in the contract-theoretic sense of the term) is crucial.

How do the data take care of (i) unobserved heterogeneity, (ii) measurement error, and (iii) reverse causality? How does the research design allow attaining the best possible identification, short of the gold standard of randomization?

Addressing these questions can only help convince reviewers of the validity of one's approach – and thus minimize the number of rejections.

Quo Vadis?

The next step is to teach our students the importance of identification.

It is puzzling that graduate students are taught about Heckman selection models before they are taught how to assess the validity of an instrumental variable.

Krugman: "The economics profession went astray because economists, as a group, mistook beauty, clad in impressive-looking mathematics, for truth."

Whither Applied Economics?

For a field that has recently rebranded itself by adding "applied" to the name of its association, the applied component sometimes leaves some to be desired.

This is all the more disheartening given that agricultural and applied economists were often at the forefront of identification.

In his 1961 article in the *Journal of Farm Economics*, for example, Mundlak defined the concept of fixed effects in an effort to get rid of the bias caused by unobserved managerial ability among farmers.

Simple Solutions

The solution lies in teaching ourselves and our students the skills required for the identification of causal relationships.

In the development seminar I teach in the fall, I tell students that one of the most important components of critical thinking is the ability to not mistake correlation for causation.

Most of the students appear get this in spite of – or is it because of? – never having taken any econometrics.

A Simple Solution

Now, I am *not* advocating that causal identification be the guiding factor behind our research. It is not a good idea for a drunk to look for his car keys under the streetlight. (It is not a good idea to drive drunk anyway!)

In other words, some important research questions are bound to remain imperfectly identified.

What I am advocating instead is for researchers to ensure that they have done due diligence when it comes to identification. This also means that most, if not all, empirical studies should discuss their identification strategy at length and acknowledge its limitations.