Journal of Economic Perspectives—Volume 24, Number 2—Spring 2010—Pages 59–68

But Economics Is Not an Experimental Science

Christopher A. Sims

Without apparent irony, Angrist and Pischke (this issue) quote Griliches (1986): "If the data were perfect, collected from well-designed randomized experiments, there would hardly be room for a separate field of econometrics." The fact is, economics is not an experimental science and cannot be. "Natural" experiments and "quasi" experiments are not in fact experiments, any more than are Prescott's "computational" experiments (for example, Kydland and Prescott, 1996). They are rhetorical devices that are often invoked to avoid having to confront real econometric difficulties. Natural, quasi-, and computational experiments, as well as regression discontinuity design, can all, when well applied, be useful, but none are panaceas. This essay by Angrist and Pischke, in its enthusiasm for some real accomplishments in certain subfields of economics, makes overbroad claims for its favored methodologies. What the essay says about macroeconomics is mainly nonsense.

The fact that the essay is so mistaken about macroeconomics reflects a broader problem. Recent enthusiasm for single-equation, linear, instrumental variables approaches in applied microeconomics has led many in these fields to avoid undertaking research that would require them to think formally and carefully about the central issues of nonexperimental inference—what Griliches saw, and I see, as the core of econometrics. Providing empirically grounded policy advice necessarily involves confronting these difficult central issues. If applied economists narrow the focus of their research and critical reading to various forms of pseudo-experimental analysis, the profession loses a good part of its ability to provide advice about the effects and uncertainties surrounding policy issues.

■ *Christopher A. Sims is the Harold H. Helm Professor of Economics and Banking, Princeton University, Princeton, New Jersey. His e-mail address is ⟨sims@princeton.edu⟩.*

doi=10.1257/jep.24.2.59

The Big Picture

Because economics is not an experimental science, economists face difficult problems of inference. The same data generally are subject to multiple interpretations. It is not that we learn nothing from data, but that we have at best the ability to use data to narrow the range of substantive disagreement. We are always combining the objective information in the data with judgment, opinion and/or prejudice to reach conclusions. Doing this well can require technically complex modeling. Doing it in a scientific spirit requires recognizing and taking account of the range of opinions about the subject matter that may exist in one's audience. That is, it requires balancing the need to use restrictive assumptions on which there may be substantial agreement against the need to leave lightly restricted those aspects of the model on which the data might help resolve disagreement.

But there are limits on the supply of able, technically tooled-up econometricians. Applied work therefore sometimes imitates the procedures of prominent, influential papers in contexts where those procedures are questionable.

The audience for applied work includes people whose interests or ideologies are affected by the outcome, but who have little technical training. There is therefore a payoff to making the methods and messages of applied work simple and easily understood, even when this involves otherwise unnecessary simplification or distortion. On the other hand, there is a danger that procedures not understood by much of the audience for a paper may lend unjustified weight to the paper's conclusions.

These tensions and pathologies have manifested themselves in different ways at different times. The Ehrlich work on capital punishment discussed at some length in the Angrist-Pischke paper is a good example. I read that work with interest at the time it appeared because it drew provocative conclusions from a new and (for the time) relatively sophisticated econometric analysis. I also discussed it with some economists at Minnesota who were preparing a critical response. What made me most uncomfortable about the paper's analysis was that it assumed a list of exogenous variables without discussing in any detail why they were both plausibly exogenous and, probably more important in that case, why the pattern of exclusion restrictions on those variables was reasonable. In fact, the only complete listing of what was assumed exogenous or predetermined appeared in footnotes to a table. The paper also implicitly invoked the idea that lagging variables made them more likely to be good instruments, which of course is not generally correct. So we were asked to believe as an a priori restriction, for example, that unemployment a year ago had an effect on this year's murder rate only via an effect on the endogenous deterrence variables, while current unemployment had a direct effect on this year's murder rate. But using instrumental variable formulas while simply listing the instruments, with little or no discussion of what kind of larger multivariate system would justify isolating the single equation or small system to which the formulas are applied, was, and to some extent still is, a common practice. Referees insisting on a more elaborate modeling framework, which no doubt would have led to mixed

conclusions rather than the provocative strong conclusion of Ehrlich's work, could easily have been seen as pedantic. Critical commentary focusing on a matter editors and referees had acquiesced in relegating to a footnote of a table might well have had difficulty getting published.

Ehrlich's work also had an element of technical hoodwinkery in the sense that its use of instrumental variables was more sophisticated than most applied microeconometrics at the time. Its stark conclusions on an ideologically charged subject attracted tremendous attention from the profession and from policymakers. And it employed a common, simplifying shortcut (listing instruments without much discussion) that was widely accepted mainly because it was widely accepted, not because it was clearly appropriate in the paper's context.

It is true that applied microeconomists these days often discuss their choices of instruments more prominently than Ehrlich did in his papers from the mid-1970s, and this is a good thing. They also have a variety of more sophisticated procedures available in packaged regression programs, like clustered standard errors. But the applied microeconomic work remains just as subject to tensions and pathologies. The Donohue and Wolfers (2005) paper that Angrist and Pischke cite as a more recent and better treatment of the subject is in good part devoted to detailed criticism of recent studies of the deterrent effect of capital punishment. The criticized studies use large modern data sets and many modern methods, including "natural experiment" language. Yet Donohue and Wolfers argue convincingly that these more recent studies are as flawed as Ehrlich's results were. Among other checks on results, Donohue and Wolfers test over-identifying restrictions on models estimated by instrumental variables, verifying that results are highly sensitive to the instrument list, and that the instruments are not plausibly all predetermined. Ehrlich could have performed this test; he used 12 instruments with three included endogenous variables, and so he had a heavily over-identified model. The test was well known at the time and easily implemented. That the more recent papers criticized by Donohue and Wolfers still failed to implement this type of test and nevertheless drew attention from policymakers is a measure of our lack of progress. Any econometric procedure used in prominent applied publications, especially if it is easy to apply, will be imitated, widely used, and eventually misused. The fact that "natural experiment" formal methods are used in questionable ways should not be a surprise. Taking the "con" out of econometrics will not be accomplished by our finding some simple, bulletproof set of methods that allow us to avoid the complexities of nonexperimental inference.

My own reaction to the Donohue and Wolfers review is that they make it clear that the murder rate varies greatly and that most of the variation is unlikely to be related to the execution rate, yet neither they nor the papers they discuss pay attention to modeling all this variation. They argue that this variation swamps death penalty deterrence effects and suggest that this makes estimating those effects hopeless. This may be true, but I would like to see a serious attempt at modeling the dynamic interactions among murder rates, policing inputs, judicial and jury choices about punishment, economic variables, drug prices and prevalence,

and other factors. Something like such a model must be used informally by any policymaker who has to make decisions about allocating resources to crime prevention. Of course this would require estimating multivariate time-series models on panel data—something for which there is no push-button in Stata. But that is where this literature ought to head. As things stand, we do not even have a good reduced-form model from which to start.

The best we can hope for is that econometricians are trained to confront the complexities and ambiguities that inevitably arise in nonexperimental inference. They should be able to fit loosely interpreted models that characterize patterns in the data, to impose identifying restrictions that allow richer interpretations, to compare the fit of alternative sets of restrictions, and to describe uncertainty about results, both in parameter estimates and across models. Angrist and Pischke might even agree with this approach in principle, but by promoting single-equation, linear, single-instrument modeling focused on single parameters and on conditional first moments alone, they are helping create an environment in which applied economists emerge from Ph.D. programs not knowing how to undertake deeper and more useful analyses of the data.

What's Taken Some of the Con out of Macroeconometrics

The Angrist and Pischke essay does not mention what seems to me the main advance in macroeconometrics: the interaction of vector autoregressions, structural vector autoregressions, and econometrically estimated dynamic stochastic general equilibrium models, which has led to broad consensus on, for example, the consequences of shifts in central bank interest rate policy.

The process began with the publication of the *Monetary History of the United States* (Friedman and Schwartz, 1963). As Rockoff (2000) points out in his review, the book was rhetorically effective, in good part because it argued based on historical "natural experiments" in which monetary quantities moved in parallel with prices and because it could be argued based on specific historical circumstances that the variation in the monetary quantities was causally prior to the inflation. There were at the time "old Keynesian" economists who believed monetary policy to be unimportant, and the Friedman and Schwartz book made that position unsustainable. But it has taken monetary economics and monetary policy decades to recover from the oversimplified message that emerged so persuasively from the Friedman and Schwartz book.

Friedman himself, as well as many other economists, argued that even in normal times the direction of causation in the correlation between money and both real and nominal variables was mainly from money supply to income. For a while in the 1970s, it was common to estimate single equations or small systems in which some measure of the money stock was treated as exogenous and to base policy conclusions on such models. I showed that in simple bivariate systems relating money and income, money did satisfy necessary conditions for exogeneity (Sims, 1972),

but later Mehra (1978) and I (Sims, 1980) showed that this conclusion broke down in systems that included an interest rate. These results implied that there was no exogenous policy variable (like "M" in the previous monetarist econometric work) that could be put on the right-hand side of a single regression equation to estimate the effects of policy. The monetary structural vector autoregression literature was a response to this fact.

The modern view among most monetary economists is that at least since 1950, and probably well before that, most variation in U.S. monetary policy has represented systematic, predictable response by the Federal Reserve to the state of the economy. As a result, estimation of the effects of monetary policy on the economy faces serious identification problems. Because most changes in the variable central banks control—a policy interest rate in nearly every case—consists of systematic response, separating the effects of monetary policy from the effects of the nonpolicy disturbances to which the central bank is responding is difficult. Romer and Romer (1989), cited favorably by Angrist and Pischke, fail to recognize this central point. They examined the record of the minutes of the Open Market Committee and identified periods when policy actions were taken *because of perceived inflationary threats.* Inflationary threats are a reflection of disturbances to the economy. There is no way to know whether the output declines that Romer and Romer estimate after policy actions are the effects of the policy actions themselves or instead are the effects of the economic disturbances that led the Open Market Committee to perceive an inflationary threat in the first place.

At an early stage, attempts to separate the effects of monetary policy on the private sector from reactions of monetary policy to the private sector in multiple equation systems brought out the "price puzzle": When identification is weak or incorrect, the equation describing monetary policy behavior tends to be confounded with the "Fisher equation," $r_t = \rho_t + E_t \pi_{t+1}$, that is, with the normal tendency of nominal interest rates to rise farther above the real rate when expected inflation is higher. When this confounding occurs, supposed contractionary monetary policy shocks are mistakenly estimated to imply higher, not lower, future inflation. Estimated systems showing a price puzzle tend to show larger real effects of monetary policy. This should not be surprising, since they are likely to confound monetary policy shocks with, for example, negative supply shocks. A negative supply shock would tend to raise interest rates as people attempt to smooth their consumption paths, and to raise prices and lower output because of the direct effects of reduced supply. It has therefore been a standard check on the accuracy of identification in these models that estimated effects of contractionary monetary policy shocks should include reduced inflation. Romer and Romer (1989) examined only the behavior of real variables in the wake of their contractionary policy dates. Leeper (1997) showed that the dummy variables Romer and Romer generate from their dates are predictable from past data, and that their unpredictable components do not behave like monetary policy shocks.

In a structural vector autoregression, restrictions based on substantive economic reasoning are placed on a multivariate time series model that allows

interpretation of some functions of its parameters as policy effects. Practitioners of this method impose identifying restrictions parsimoniously, often leaving most parameters of the model without a behavioral interpretation, and usually leaving the fit of the model to the multivariate time series data as good as that of an unrestricted reduced form model. In monetary policy vector autoregressions, a variety of restrictions have appeared in the literature. Some assume that there is a delay of a month or a quarter between an interest rate change and its effects on output or on consumer prices (but not on asset or commodity prices). Others assume that the response to a monetary policy tightening must produce a fall, or no change, in output and prices, or that a monetary policy change can have no long-run effect on real variables. With this variety of identifying assumptions, a consistent picture has emerged: monetary contraction produces a decline in output and a decline in inflation, with both responses smooth and delayed and the decline in output quicker. Of course, the sign pattern of these responses was used in identification, formally in some cases and informally in others, but the results are quantitatively, not just qualitatively, consistent across identification strategies. There are two robust results that were not entailed by the identifying assumptions: First, if one believes monetary contraction immediately raises interest rates and then is followed by decreases (or no change) in output or inflation, there is evidence in the data of some random variation in monetary policy fitting that pattern. Second, it is not possible to attribute more than a small fraction of cyclical variation in output or interest rates to such random variation in monetary policy. There is also evidence that switching to a monetary policy rule that reacted less strongly to inflation than the historically observed rule (or even that merely followed Friedman's prescription of stabilizing the growth rate of the money stock) would have resulted in a more volatile time path for inflation than was historically observed (Sims and Zha, 2006).

In the last few years, starting with the work of Smets and Wouters (2003), models with more complete interpretations than the structural vector autoregressions, which fit nearly as well as structural vector autoregressions, have been estimated. These models, called dynamic stochastic general equilibrium models, make much stronger assumptions than the structural vector autoregressions, but they reproduce the implications of the structural vector autogressions for the effects of monetary policy. The fact that the models match in this respect increases confidence that the dynamic stochastic general equilibrium models are not getting their estimates of monetary policy effects mainly from their strong assumptions. The dynamic stochastic general equilibrium models have the advantage as a framework for policy discussion that they make explicit why the effects of policy take the form they do and that they allow interpretation of the sources of nonpolicy disturbances to the economy. These models are for the most part estimated by treating the shape of the likelihood as characterizing the uncertainty about parameters—that is, by taking a Bayesian perspective on inference. This is what makes it possible to describe uncertainty about the implications of both kinds of models in a consistent framework that accounts for parameter uncertainty, and thereby to combine uncertain judgmental information with model results.

There are other interesting developments in macroeconometrics, but for the purposes of this comment on Angrist and Pischke, the narrative above makes my point. In macro, the turning away from mechanical imposition of zero restrictions and exogeneity assumptions in the simultaneous equations literature led to vector autogressions and structural vector autoregressions. In these models, identifying assumptions are relatively few and are at the center in presenting results, which in a limited way makes these methods similar to the "design-based inference" approach that Angrist and Pischke endorse. But the developments in empirical macroeconomics have been different in important dimensions. The models are still multiple-equation. Attempts to use "natural experiment" language to justify particular identifying assumptions have not been very successful or very influential. And the parsimoniously identified models are being systematically related to more heavily restricted and completely interpreted models that can be used for actual policy analysis.

Multiple-Equation Models, Nonlinear Models, Generalized Least Squares, Mixed and Mixture Models

The question of how class size affects educational achievement receives a lot of attention in the Angrist and Pishke essay. The cumulative impact of the work on this issue cited in the essay is undoubtedly a success story for the methodologies the essay pushes. But what question has been answered? Who is to use the result, and for what?

The implicit audience here is educational policymakers. If I were a school principal looking at the results from a regression discontinuity design study, my first question would be, how are the reductions in class size when an enrollment threshold is crossed in these studies being achieved? Does the principal get sent additional teachers with experience teaching the relevant grade level, drawn from other schools in the system with lower enrollment in that grade? Or does the principal have to adjust resources within that principal's own school subject to a fixed budget of teacher count and/or dollars? Or (and this is surely unlikely) does the school system hire new, experienced, teachers whenever a particular school and grade hits the enrollment limit?

As the essay points out, we know that principals tend to put the lowest-achieving or most disruptive students in small classes, presumably with some objective in mind. If I were a school system administrator, I would want to know whether imposing a rule on my principals that they must not have any classes larger than, say, 25 would be good policy. This policy would obviously limit the principals' ability to adjust class sizes according to the criteria they are currently using, unless I promised to provide additional teachers and space whenever a class reached an enrollment of 25.

The results from Tennessee's Student Teacher Achievement Ratio (STAR) experimental study make it clearer how the differences in class sizes are being

achieved, but the source of variation here is related only to a narrow, and perhaps not very interesting, range of feasible policy actions. Real world policy actions seem unlikely to take the form of hiring new teachers and building new classrooms in order uniformly to reduce the size of all classes. They instead seem most likely to result in changes in the distribution of class sizes. Would reduced inequality of class sizes, as might emerge from an upper limit on class size, be a good thing? The STAR study might actually go some way to answering that question, though in the discussion I have seen of it, the emphasis seems to be entirely on whether the students in the smaller classes are better off, not on whether reducing inequality of class sizes would be an improvement. Here a careful exploration of possible nonlinearities would be of central importance. Do the effects of larger class sizes taper off above some size, or do they increase steeply? Do the effects of smaller sizes drop off below some small class size? The linear instrumental variable estimates of average effects (with robust standard errors corrected for heteroskedasticity and autocorrelation, of course), with which Angrist and Pischke seem happy, are inherently inadequate to answer such real policy questions.

Angrist and Pischke argue that worrying about nonlinearity and about modeling error distributions is a "distraction," endorsing the increasingly common practice of using linear models combined with robust standard errors. This approach is justifiable in a regression model, but only if one takes the view that 1) we are interested in estimating the coefficients of the best linear predictor of y based on X; and 2) we believe that $\mathrm{E}[y \mid X]$ is *not* linear, so it is important to recognize the part of the error due to misspecification in the linear model. (See Szpiro, Rice, and Lumley (2008) and Chamberlain (1987) for elaboration of this point.) In that case, ordinary least squares regressions with robust standard errors is close to the best we can do. But this is a case where the coefficients of the linear model being estimated depend on the distribution of the right-hand-side variables. There are few applications where that kind of a model is really of interest. We usually are aiming at estimating $\mathrm{E}[y \mid X]$ accurately. In that case, if linearity is a good approximation, weighted or generalized least squares gives a clearer picture of what is going on in the data. In many applications—both the class size and capital punishment models, for example—there is a lot of interest in whether estimates are significantly different from zero. The use of generalized least squares can make a huge difference to conclusions in such cases. Even better, one can use what statisticians call "mixed models," in which conditional heteroskedasticity is modeled as arising from random variation in coefficients. Instead of clustering standard errors by state in a state panel data application, for example, one would model coefficients as varying randomly across states. With this approach, unlike with clustered standard errors, one can gain insight into the nature of conditional heteroskedasticity and thereby into the nature of heterogeneity across states. These models are straightforward to handle with modern Bayesian, Markov-chain Monte Carlo methods. To explore simultaneously for nonlinearity and nonscalar covariance of residuals, an easily implemented approach is laid out in Norets (2009). Observing that robust standard errors are quite different from conventional ones, which do not cluster or account

for heteroskedasticity, should be a signal to us that there is a great deal going on in the data that our linear model is missing. Accounting for it carefully could change our conclusions about the strength of evidence on linear effects and might also lead us to question whether the linear model is answering the most interesting questions about the data.

Conclusion

Natural experiments, difference-in-difference, and regression discontinuity design are good ideas. They have not taken the con out of econometrics—in fact, as with any popular econometric technique, they in some cases have become the vector by which "con" is introduced into applied studies. Furthermore, overenthusiasm about these methods, when it leads to claims that single-equation linear models with robust standard errors are all we ever really need, can lead to our training applied economists who do not understand fully how to model a dataset. This is especially regrettable because increased computing power—and the new methods of inference that are arising to take advantage of this power—make such narrow, overly simplified approaches to data analysis increasingly obsolete.

References

Chamberlain, Gary. 1987. "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions." *Journal of Econometrics*, 34(3): 305–34.

Donohue, John J., and Justin Wolfers. 2005. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate." *Stanford Law Review*, vol. 58, pp. 791–845.

Friedman, Milton, and Anna. J. Schwartz. 1963. *A Monetary History of the United States, 1867–1960.* Princeton University Press.

Griliches, Zvi. 1986. "Economic Data Issues." In *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. 3, pp. 1465–1514. North-Holland, Amsterdam.

Kydland, Finn E., and Edward C. Prescott. 1996. "The Computational Experiment: An Econometric Tool." *Journal of Economic Perspectives*, 10(1): 69–85.

Leeper, Eric M. 1997. "Narrative and VAR Approaches to Monetary Policy: Common Identification Problems." *Journal of Monetary Economics*, 40(3): 641–58.

Mehra, Yash P. 1978. "Is Money Exogenous in Money-Demand Equations." *Journal of Political Economy*, 86(2): 211–28.

Norets, Andriy. 2009. "Approximation of Conditional Densities by Smooth Mixtures of Regressions." http://www.princeton.edu/~anorets/mixreg.pdf.

Rockoff, Hugh. 2000. Review of *A Monetary History of the United States, 1867–1960*, by Milton Friedman and Anna Jacobson Schwartz. EH.Net Economic History Services. http://eh.net/bookreviews/library/rockoff.

Romer, Christina D., and David H. Romer. 1989. "Does Monetary Policy Matter?: A New Test in the Spirit of Friedman and Schwartz." *NBER Macroeconomics Annual*, vol. 4, 121–170.

Sims, Christopher A. 1972. "Money, Income, and Causality." *American Economic Review*, 62(4): 540–52.

Sims, Christopher A. 1980. "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered." *American Economic Review*,

70(2): 250–57.

Sims, Christopher A., and Tao Zha. 2006. "Does Monetary Policy Generate Recessions?" *Macroeconomic Dynamics*, 10(2): 231–72.

Smets, Frank, and Raf Wouters. 2003. "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area." *Journal of the European Economic Association*, 1(5): 1123–75.

Szpiro, Adam A., Kenneth M. Rice, and Thomas Lumley. 2008. "Model-Robust Regression and a Bayesian 'Sandwich' Estimator." UW Biostatistics Working Paper 338, University of Washington University. http://www.bepress.com/uwbiostat/paper338.

This article has been cited by:

1. G. Andrew Karolyi. 2011. The Ultimate Irrelevance Proposition in Finance?. *Financial Review* **46**:4, 485-512. [CrossRef]
2. François Claveau. 2011. Evidential variety as a source of credibility for causal inference: beyond sharp designs and structural models. *Journal of Economic Methodology* **18**:3, 233-253. [CrossRef]
3. Angus Deaton. 2010. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature* **48**:2, 424-455. [Abstract] [View PDF article] [PDF with links]
4. James J. Heckman. 2010. Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy. *Journal of Economic Literature* **48**:2, 356-398. [Abstract] [View PDF article] [PDF with links]