COMMENT ON ANGRIST AND PISCHKE

CHRISTOPHER A. SIMS

I. The bottom line, at the top

Without apparent irony, Angrist and Pischke quote Griliches: "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. They are rhetorical devices invoked to avoid having to confront real econometric difficulties. Natural, quasi-, and computational experiments, as well as regression discontinuity design (RDD), 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 to many economists in these fields losing the ability to think formally and carefully about the central issues of non-experimental inference — what Griliches saw, and I see, as the core of econometrics.

II. The big picture

Because we are not an experimental science, we 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 your audience.

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.

Date: January 4, 2010.

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. And on the other hand there is room for procedures that are not understood by much of the audience for a paper to 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 and discussed it with some economists at Minnesota who were preparing a critical response. It was clear to me then that the main problem with the paper 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 this case, why the pattern of exclusion restrictions on those variables was reasonable. In fact the only complete listing of what it assumed exogenous or predetermined was in the footnotes to a table. It 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, but only via an effect on the endogenous deterrence variables, while current unemployment had a direct impact. But using instrumental variables 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 who insisted on a more elaborate modeling framework, which would no doubt have led to mixed conclusions rather than the provocative strong conclusion of Ehlrich's work, might easily have been seen as pedantic. And critical commentary that emphasized what the editors and referees had acquiesced in relegating to a table footnote might well have had difficulty getting published.

So Ehrlich's work had an element of technical hoodwinkery — its use of instrumental variables was more sophisticated than most applied microeconometrics at the time. Its stark conclusions on an ideologically charged subject gave it a tremendous boost in attention from the profession and from policy-makers. 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, and this is a good thing. They also

have available a variety of more sophisticated procedures available in packaged regression programs, like clustered standard errors. But the applied microeconomic work remains fully as subject to tensions and pathologies. The Donohue and Wolfers paper that AP 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 they are as flawed as were Ehrlich's results. 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, so had a heavily overidentified model. The test was well known at the time and easily implemented. That the recently written papers Donohue and Wolfers criticize still failed to implement this type of test and still drew attention from policy makers is a measure of our lack of progress.

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, etc. Something like such a model has to be used informally by any policy-maker 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 that there is no push-button for in Stata. But that is where this literature ought to head. As things stand, we do not even have a good reduced-form model to start from.

The best we can hope for is that econometricians are trained to confront the complexities and ambiguities that inevitably arise in non-experimental 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. AP would probably agree with this 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 PhD programs not knowing how to undertake deeper and more useful analyses of the data.

III. What's taken some of the con out of macroeconometrics

The AP essay does not mention what seems to me the main advance in macroeconometrics. The interaction of vector autoregressions (VARs), structural vector autoregressions (SVARs) and econometrically estimated dynamic stochastic general equilibrium models (DSGEs) has led to broad consensus on 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 points out in his 2000 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 in which 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 to income. For a while in the 1970's, 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 1972 that in simple bivariate systems money did satisfy necessary conditions for exogeneity, but later Mehra (1978) and I 1980 showed that this conclusion broke down in systems that included an interest rate.

The modern view among most monetary economists is that at least since 1950, and probably long before that, most variation in US monetary policy has represented systematic, predictable response by the Fed to the state of the economy. This means that 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 non-policy disturbances to which the central bank is responding is difficult. Romer and Romer (1989), cited favorably by AP, failed entirely to recognize this central point. They examined the record of the minutes of the FOMC and identified periods when policy actions were taken *because of perceived inflationary threats*. There is no way to know whether the pattern of behavior of economic variables after such actions reflected the effects of the policy actions or instead the effects of the developments that led the FOMC to perceive an inflationary threat.

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. Unsurprisingly, estimated systems showing a price puzzle tend to show larger real effects of monetary policy, since they are likely to confound monetary policy shocks with negative supply shocks or with non-monetary-policy demand shocks that push the economy close to or above its capacity limits. 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 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.

The monetary structural VAR literature produced numerous papers, using data from different countries and on different lists of variables and using a variety of approaches to identification. A consistent picture 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. This is not really the result of the literature, though, since the signs of this pattern of responses were used as identifying restrictions, formally in some cases, informally in others. The robust results are, first, that if one believes monetary contraction immediately raises interest rates and then is followed by increases in neither output nor inflation, there is evidence in the data of some random variation in monetary policy fitting that pattern and, second, that 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 SVAR's and that fit nearly as well as SVAR's have been estimated. These models, called DSGE's (dynamic stochastic general equilibrium models) make much stronger assumptions than the SVAR's, but they reproduce the SVAR's implications for the effects of monetary policy. The fact

that the models match in this respect increases confidence that the DSGE's are not getting their estimates of monetary policy effects mainly from their strong assumptions. The DSGE's 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 non-policy 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 — i.e., taking a Bayesian perspective on inference. This is what has made it possible to compare their fit to that of the less-structured SVAR's. It also makes it possible to describe uncertainty about the models' implications 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 AP the narrative above makes my point. In macro, the turning away from mechanical imposition of zero restrictions and exogeneity assumptions led to VAR's and SVAR's. These are models in which identifying assumptions are relatively few and are at the center of presentation of results, as in the "design-based inference" approaches AP endorse. But in macro, 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.

IV. Multiple-equation models; nonlinear models; GLS; mixed and mixture models

The class size question receives a lot of attention in the 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 is the question that has been answered? Who is to use the result, and for what?

The implicit audience here is educational policy-makers. If I were a school principal looking at the RDD results, 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 his own school subject to a fixed budget of teacher count and/or dollars? Or (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, 40 would be good policy. This 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 40.

The STAR experiment results make it clearer how the differences in class sizes are being achieved, but the source of variation here is no more related to feasible policy actions than in the RDD studies. Real world policies manipulating class size seem most likely to result in reduced inequality of class sizes. Is this 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 non-linearities 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 IV estimates of average effects (with HAC standard errors of course) that AP seem happy with are inherently inadequate to answer such real policy questions.

AP argue that worrying about nonlinearity and about modeling error distributions is a "distraction", endorsing the increasingly common practice of using linear models combined with "sandwich" 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 $E[y \mid X]$ is *not* linear, so that 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 OLS with sandwich 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 $E[y \mid X]$ accurately. In that case, if linearity is a good approximation, GLS gives a clearer picture of what is going on in the data. Note that dismissing GLS as merely improving efficiency is a mistake. 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. GLS 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, e.g., 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 easy to handle with modern Bayesian MCMC methods. To explore simultaneously for nonlinearity and non-scalar covariance of residuals, an easily implemented approach is laid out in Norets (2009).

V. 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, over-enthusiasm about these methods, when it leads to claims that single-equation linear model with sandwiched errors are all we ever really need, can lead to our training applied economists who do not understand how to fully 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, simple approaches to data analysis increasingly obsolete.

References

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Friedman, M., and A. J. Schwartz (1963): *A Monetary History of the United States, 1867-1960*. Princeton University Press.

Leeper, E. (1997): "Narrative and VAR approaches to monetary policy: Common identification problems," *Journal of Monetary Economics 40, December, 1997: 641-658*, 40, 641–658.

Mehra, Y. P. (1978): "Is Money Exogenous in Money-Demand Equations," *The Journal of Political Economy*, 86(2), 211–228.

Norets, A. (2009): "Approximation of conditional densities by smooth mixtures of regressions," Discussion paper, Princeton University, `http://www.princeton.edu/~anorets/mixreg.pdf`.

Rockoff, H. (2000): "Review of Milton Friedman and Anna Jacobson Schwartz, A Monetary History of the United States, 1867-1960," EH.Net Economic History Services.

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

SIMS, C. A. (1972): "Money, Income, and Causality," *The American Economic Review*, 62(4), 540–552.

——— (1980): "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered," *American Economic Review*, 70, 250–57.

SIMS, C. A., AND T. ZHA (2006): "Does Monetary Policy Generate Recessions?," *Macroeconomic Dynamics*, 10(2), 231–272.

SMETS, F., AND R. WOUTERS (2003): "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area," *Journal of the European Economic Association*, 1, 1123–1175.

SZPIRO, A. A., K. M. RICE, AND T. LUMLEY (2008): "Model-Robust Regression and a Bayesian 'Sandwich' Estimator," UW Biostatistics Working Paper Series 338, University of Washington University, http://www.bepress.com/uwbiostat/paper338.