Journal of Economic Perspectives—Volume 24, Number 2—Spring 2010—Pages 31–46

Tantalus on the Road to Asymptopia

Edward E. Leamer

My first reaction to "The Credibility Revolution in Empirical Economics," authored by Joshua D. Angrist and Jörn-Steffen Pischke, was: Wow! This paper makes a stunningly good case for relying on purposefully randomized or accidentally randomized experiments to relieve the doubts that afflict inferences from nonexperimental data. On further reflection, I realized that I may have been overcome with irrational exuberance. Moreover, with this great honor bestowed on my "con" article, I couldn't easily throw this child of mine overboard.

We economists trudge relentlessly toward Asymptopia, where data are unlimited and estimates are consistent, where the laws of large numbers apply perfectly and where the full intricacies of the economy are completely revealed. But it's a frustrating journey, since, no matter how far we travel, Asymptopia remains infinitely far away. Worst of all, when we feel pumped up with our progress, a tectonic shift can occur, like the Panic of 2008, making it seem as though our long journey has left us disappointingly close to the State of Complete Ignorance whence we began.

The pointlessness of much of our daily activity makes us receptive when the Priests of our tribe ring the bells and announce a shortened path to Asymptopia. (Remember the Cowles Foundation offering asymptotic properties of simultaneous equations estimates and structural parameters?) We may listen, but we don't hear, when the Priests warn that the new direction is only for those with Faith, those with complete belief in the Assumptions of the Path. It often takes years down the Path, but sooner or later, someone articulates the concerns that gnaw away in each of

■ *Edward E. Leamer is Professor of Economics, Management and Statistics, University of California at Los Angeles, Los Angeles, California. His e-mail address is ⟨edward.leamer@anderson.ucla.edu⟩.*

doi=10.1257/jep.24.2.31

us and asks if the Assumptions are valid. (T. C. Liu (1960) and Christopher Sims (1980) were the ones who proclaimed that the Cowles Emperor had no clothes.) Small seeds of doubt in each of us inevitably turn to despair and we abandon that direction and seek another.

Two of the latest products-to-end-all-suffering are nonparametric estimation and consistent standard errors, which promise results without assumptions, as if we were already in Asymptopia where data are so plentiful that no assumptions are needed. But like procedures that rely explicitly on assumptions, these new methods work well in the circumstances in which explicit or hidden assumptions hold tolerably well and poorly otherwise. By disguising the assumptions on which nonparametric methods and consistent standard errors rely, the purveyors of these methods have made it impossible to have an intelligible conversation about the circumstances in which their gimmicks do not work well and ought not to be used. As for me, I prefer to carry parameters on my journey so I know where I am and where I am going, not travel stoned on the latest euphoria drug.

This is a story of Tantalus, grasping for knowledge that remains always beyond reach. In Greek mythology Tantalus was favored among all mortals by being asked to dine with the gods. But he misbehaved—some say by trying to take divine food back to the mortals, some say by inviting the gods to a dinner for which Tantalus boiled his son and served him as the main dish. Whatever the etiquette faux pas, Tantalus was punished by being immersed up to his neck in water. When he bowed his head to drink, the water drained away, and when he stretched up to eat the fruit hanging above him, wind would blow it out of reach. It would be much healthier for all of us if we could accept our fate, recognize that perfect knowledge will be forever beyond our reach and find happiness with what we have. If we stopped grasping for the apple of Asymptopia, we would discover that our pool of Tantalus is full of small but enjoyable insights and wisdom.

Can we economists agree that it is extremely hard work to squeeze truths from our data sets and what we genuinely understand will remain uncomfortably limited? We need words in our methodological vocabulary to express the limits. We need sensitivity analyses to make those limits transparent. Those who think otherwise should be required to wear a scarlet-letter *O* around their necks, for "overconfidence." Angrist and Pischke obviously know this. Their paper is peppered with concerns about quasi-experiments and with criticisms of instrumental variables thoughtlessly chosen. I think we would make progress if we stopped using the words "instrumental variables" and used instead "surrogates"—meaning surrogates for the experiment that we wish we could have conducted. The psychological power of the vocabulary requires a "surrogate" to be chosen with much greater care than an "instrument."

As Angrist and Pischke persuasively argue, either purposefully randomized experiments or accidentally randomized "natural" experiments can be extremely helpful, but Angrist and Pischke seem to me to overstate the potential benefits of the approach. Since hard and inconclusive thought is needed to transfer the

results learned from randomized experiments into other domains, there must therefore remain uncertainty and ambiguity about the breadth of application of any findings from randomized experiments. For example, how does Card's (1990) study of the effect on the Miami labor market of the Mariel boatlift of 125,000 Cuban refugees in 1980 inform us of the effects of a 2000 mile wall along the southern border of the United States? Thoughts are also needed to justify the choice of instrumental variables, and a critical element of doubt and ambiguity necessarily afflicts any instrumental variables estimate. (You and I know that truly consistent estimators are imagined, not real.) Angrist and Pischke understand this. But their students and their students' students may come to think that it is enough to wave a clove of garlic and chant "randomization" to solve all our problems just as an earlier cohort of econometricians have acted as if it were enough to chant "instrumental variable."

I will begin this comment with some thoughts about the inevitable limits of randomization, and the need for sensitivity analysis in this area, as in all areas of applied empirical work. To be provocative, I will argue here that the financial catastrophe that we have just experienced powerfully illustrates a reason why extrapolating from natural experiments will inevitably be hazardous. The misinterpretation of historical data that led rating agencies, investors, and even myself to guess that home prices would decline very little and default rates would be tolerable even in a severe recession should serve as a caution for all applied econometrics. I will also offer some thoughts about how the difficulties of applied econometric work cannot be evaded with econometric innovations, offering some under-recognized difficulties with instrumental variables and robust standard errors as examples. I conclude with some comments about the shortcomings of an experimentalist paradigm as applied to macroeconomics, and with some warnings about the willingness of applied economists to apply push-button methodologies without sufficient hard thought regarding their applicability and shortcomings.

Randomization Is Not Enough

Angrist and Pischke offer a compelling argument that randomization is one large step in the right direction. Which it is! But like all the other large steps we have already taken, this one doesn't get us where we want to be.

In addition to randomized treatments, most scientific experiments also have controls over the important confounding effects. These controls are needed to improve the accuracy of the estimate of the treatment effect and also to determine clearly the range of circumstances over which the estimate applies. (In a laboratory vacuum, we would find that a feather falls as fast as a bowling ball. In the real world with air, wind, and humidity, all bets are off, pending further study.)

In place of experimental controls, economists can, should, and usually do include control variables in their estimated equations, whether the data are

nonexperimental or experimental. To make my point about the effect of these controls it will be helpful to refer to the prototypical model:

$$y_t = \alpha + (\beta_0 + \boldsymbol{\beta}_1'\mathbf{z}_t)\, x_t + \boldsymbol{\theta}'\mathbf{w}_t + \varepsilon_t,$$

where x is the treatment, y the response, $\mathbf{z}$ is a set of interactive confounders, $\mathbf{w}$ is a set of additive confounders, where ε stands for all the other unnamed, unmeasured effects which we sheepishly assume behaves like a random variable, distributed independently of the observables. Here $(\beta_0 + \boldsymbol{\beta}_1'\mathbf{z}_t)$ is the variable treatment effect that we wish to estimate. One set of problems is caused by the additive confounding variables $\mathbf{w}$, which can be uncomfortably numerous. Another set of problems is caused by the interactive confounding variables $\mathbf{z}$, which may include features of the experimental design as well as characteristics of the subjects.

Consider first the problem of the additive confounders. We have been taught that experimental randomization of the treatment eliminates the requirement to include additive controls in the equation because the correlation between the controls and the treatment is zero by design and regression estimates with or without the controls are unbiased, indeed identical. That's true in Asymptopia, but it's not true here in the Land of the Finite Sample where correlation is an ever-present fact of life and where issues of sensitivity of conclusions to assumptions can arise even with randomized treatments if the correlations between the randomized treatment and the additive confounders, by chance, are high enough.

Indeed, if the number of additive confounding variables is equal to or larger than the number of observations, any treatment x, randomized or not, will be perfectly collinear with the confounding variables (the undersized sample problem). Then, to estimate the treatment effect, we would need to make judgments about which of the confounding variables to exclude. That would ordinarily require a sensitivity analysis, unless through Divine revelation economists were told exactly which controls to include and which to exclude. Though the number of randomized trials may be large, an important sensitivity question can still arise because the number of confounding variables can be increased without limit by using lagged values and nonlinear forms. In other words, if you cannot commit to some notion of smoothness of the functional form of the confounders and some notion of limited or smooth time delays in response, you will not be able to estimate the treatment effect even with a randomized experiment, unless experimental controls keep the confounding variables constant or Divine inspiration allows you to omit some of the variables.

You are free to dismiss the preceding paragraph as making a mountain out of a molehill. By reducing the realized correlation between the treatment and the controls, randomization allows a larger set of additive control variables to be included before we confront the sensitivity issues caused by collinearity. For that reason, though correlation between the treatment and the confounders with nonexperimental data is a *huge* problem, it is much less important when the treatment is randomized. With that problem neutralized, concern shifts elsewhere.

The big problem with randomized experiments is not additive confounders; it's the interactive confounders. This is the heterogeneity issue that especially concerns Heckman (1992) and Deaton (2008) who emphasized the need to study "causal mechanisms," which I am summarizing in terms of the interactive **z** variables. Angrist and Pischke completely understand this point, but they seem inappropriately dismissive when they accurately explain "extrapolation of causal effects to new settings is always speculative," which is true, but the extrapolation speculation is more transparent and more worrisome in the experimental case than in the nonexperimental case.

After all, in nonexperimental nonrandomized settings, when judicious choice of additive confounders allows one to obtain just about any estimate of the treatment effect, there is little reason to worry about "extrapolation of causal effects to new settings." What's to extrapolate anyway? Our lack of knowledge? Greater concern about extrapolation is thus an indicator of the progress that comes from randomization.

When the randomization is accidental, we may pretend that the instrumental variables estimator is consistent, but we all know that the assumptions that justify that conclusion cannot possibly hold exactly. Those who use instrumental variables would do well to anticipate the inevitable barrage of questions about the appropriateness of their instruments. Ever-present asymptotic bias casts a large dark shadow on instrumental variables estimates and is what limits the applicability of the estimate even to the setting that is observed, not to mention extrapolation to new settings. In addition, small sample bias of instrumental variables estimators, even in the consistent case, is a huge neglected problem with practice, made worse by the existence of multiple weak instruments. This seems to be one of the points of the Angrist and Pischke paper—purposeful randomization is better than accidental randomization.

But when the randomization is purposeful, a whole new set of issues arises—experimental contamination—which is much more serious with human subjects in a social system than with chemicals mixed in beakers or parts assembled into mechanical structures. Anyone who designs an experiment in economics would do well to anticipate the inevitable barrage of questions regarding the valid transference of things learned in the lab (one value of z) into the real world (a different value of z).

With interactive confounders explicitly included, the overall treatment effect $\beta_0 + \boldsymbol{\beta}'\mathbf{z}_t$ is not a number but a variable that depends on the confounding effects. Absent observation of the interactive compounding effects **z**, what is estimated is some kind of average treatment effect which is called by Imbens and Angrist (1994) a "Local Average Treatment Effect," which is a little like the lawyer who explained that when he was a young man he lost many cases he should have won but as he grew older he won many that he should have lost, so that on the average justice was done. In other words, if you act as if the treatment effect is a random variable by substituting β_t for $\beta_0 + \boldsymbol{\beta}'\mathbf{z}_t$, the notation inappropriately relieves you of the heavy burden of considering what are the interactive confounders and finding some way to measure them. Less elliptically, absent observation of **z**, the estimated treatment

effect should be transferred *only* into those settings in which the confounding interactive variables have values close to the mean values in the experiment. If little thought has gone into identifying these possible confounders, it seems probable that little thought will be given to the limited applicability of the results in other settings. This is the error made by the bond rating agencies in the recent financial crash—they transferred findings from one historical experience to a domain in which they no longer applied because, I will suggest, social confounders were not included. More on this below.

Sensitivity Analysis and Sensitivity Conversations are What We Need

I thus stand by the view in my 1983 essay that econometric theory promises more than it can deliver, because it requires a complete commitment to assumptions that are actually only half-heartedly maintained. The only way to create credible inferences with doubtful assumptions is to perform a sensitivity analysis that separates the fragile inferences from the sturdy ones: those that depend substantially on the doubtful assumptions and those that do not. Since I wrote my "con in econometrics" challenge much progress has been made in economic theory and in econometric theory and in experimental design, but there has been little progress technically or procedurally on this subject of sensitivity analyses in econometrics. Most authors still support their conclusions with the results implied by several models, and they leave the rest of us wondering how hard they had to work to find their favorite outcomes and how sure we have to be about the instrumental variables assumptions with accidentally randomized treatments and about the extent of the experimental bias with purposefully randomized treatments. It's like a court of law in which we hear only the experts on the plaintiff's side, but are wise enough to know that there are abundant arguments for the defense.

I have been making this point in the econometrics sphere since before I wrote *Specification Searches: Ad Hoc Inference with Nonexperimental Data* in 1978.[1] That book was stimulated by my observation of economists at work who routinely pass their data through the filters of many models and then choose a few results for reporting purposes. The range of models economists are willing to explore creates ambiguity in the inferences that can properly be drawn from our data, and I have been recommending mathematical methods of sensitivity analysis that are intended to determine the limits of that ambiguity.

[1] Parenthetically, if you are alert, you might have been unsettled by the use of the word "with" in my title: *Ad Hoc Inference with Nonexperimental Data,* since inferences are made *with tools* but *from data.* That is my very subtle way of suggesting that knowledge is created by an interactive exploratory process, quite unlike the preprogrammed estimation dictated by traditional econometric theory.

The language that I used to make the case for sensitivity analysis seems not to have penetrated the consciousness of economists. What I called "extreme bounds analysis" in my 1983 essay is a simple example that is the best-known approach, though poorly understood and inappropriately applied. Extreme bounds analysis is not an "ad hoc but intuitive approach," as described by Angrist and Pischke. It is a solution to a clearly and precisely defined sensitivity question, which is to determine the range of estimates that the data could support given a precisely defined range of assumptions about the prior distribution. It's a correspondence between the assumption space and the estimation space. Incidentally, if you could see the wisdom in finding the range of estimates that the data allow, I would work to provide tools that identify the range of *t*-values, a more important measure of the fragility of the inferences.

A prior distribution is a foreign concept for most economists, and I tried to create a bridge between the logic of the analysis and its application by expressing the bounds in language that most economists could understand. Here it is: Include in the equation the treatment variable and a single linear combination of the additive controls. Then find the linear combination of controls that provides the greatest estimated treatment effect and the linear combination that provides the smallest estimated treatment effect. That corresponds to the range of estimates that can be obtained when it is known that the controls are doubtful (zero being the most likely estimate, a priori) but there is complete ambiguity about the probable importance of the variables, arbitrary scales, and arbitrary coordinate systems. What is ad hoc are the follow-on methods, for example, computing the standard errors of the bounds or reporting a distribution of estimates as in Sala-i-Martin (1997).

A culture that insists on statistically significant estimates is not naturally receptive to another reason our data are uninformative (too much dependence on arbitrary assumptions). One reason these methods are rarely used is their honesty seems destructive; or, to put it another way, a fanatical commitment to fanciful formal models is often needed to create the appearance of progress. But we need to change the culture and regard the finding of "no persuasive evidence in these data" on the same footing as a "statistically significant and sturdy estimate." Keep in mind that the sensitivity correspondence between assumptions and inferences can go in either direction. We can ask what set of inferences corresponds to a particular set of assumptions, but we can also ask what assumptions are needed to support a hoped-for inference. That is exactly what an economic theorem does. The intellectual value of the Factor Price Equalization Theorem does not derive from its truthfulness; it's value comes from the fact that it provides a minimal set of assumptions that imply Factor Price Equalization, thus focusing attention on *why* factor prices are not equalized.

To those of you who do data analysis, I thus pose two questions that I think every empirical enterprise should be able to answer: What feature of the data leads to that conclusion? What set of assumptions is essential to support that inference?

The Troubles on Wall Street and Three-Valued Logic

With the ashes of the mathematical models used to rate mortgage-backed securities still smoldering on Wall Street, now is an ideal time to revisit the sensitivity issues. Justin Fox (2009) suggests in his title *The Myth of the Rational Market, A History of Risk, Reward, and Delusion on Wall Street* that we have been making a modeling error and that the problems lie in the assumption of rational actors, which presumably can be remedied by business-as-usual after adding a "behavioral" variable or two into the model. I think the roots of the problem are deeper, calling for a change in the way we do business and calling for a book that might be titled: *The Myth of the Data Generating Process: A History of Delusion in Academia*. Rationality of financial markets is a pretty straightforward consequence of the assumption that financial returns are drawn from a "data generating process" whose properties are apparent to experienced investors and econometricians, after studying the historical data. If we don't know the data generating process, then the efficient markets edifice falls apart. Even simple-minded finance ideas, like the benefits from diversification, become suspect if we cannot reliably assess predictive means, variances, and covariances.

But it isn't just finance that rests on this myth of a data-generating process. It is the whole edifice of empirical economics. Let's face it. The evolving, innovating, self-organizing, self-healing human system we call the economy is not well described by a fictional "data-generating process." The point of the sensitivity analyses that I have been advocating begins with the admission that the historical data are compatible with countless alternative data-generating models. If there is one, the best we can do is to get close; we are never going to know it.

Confronted with our collective colossal failure to anticipate the problems with mortgage-backed securities, we economists have been stampeding shamelessly back to Keynesian thinking about macroeconomics, scurrying to reread Keynes' (1936) *General Theory of Employment, Interest and Money*. We would do well to go back a little further in time to 1921 when both Keynes' *Treatise on Probability* and Frank Knight's *Risk, Uncertainty and Profit* were published. Both of these books are about the myth of the data generating process. Both deal with the limits of the expected utility maximization paradigm. Both serve as foundations for the arguments in favor of sensitivity analysis in my "con" paper. Both are about "three-valued logic."

Here is how three-valued logic works. Suppose you can confidently determine that it is a good idea to bring your umbrella if the chance of rain is 10 percent or higher, but carrying the umbrella involves more cost than expected benefit if the chance of rain is less. When the data point clearly to a probability either in excess of 10 percent or less than 10 percent, then we are in a world of Knightian risk in which the decision can be based on expected utility maximization. But suppose there is one model that suggests the probability is 15 percent while another equally good model suggests the probability is 5 percent. Then we are in a world of Knightian

uncertainty in which expected utility maximization doesn't produce a decision. When any number between 0.05 and 0.15 is an equally good[2] assessment of the chance of rain we are dealing with epistemic probabilities that are not numbers but intervals, in the spirit of Keynes (1921). While the decision is two-valued—you either take your umbrella or you don't—the state of mind is three-valued: yes (take the umbrella), no (leave it behind), or I don't know. The point of adopting three-valued logic like Keynes and Knight is to encourage us to think clearly about the limits of our knowledge and the limits of the expected utility maximization paradigm. With all of the focus on decision making with two-valued logic, the profession has done precious little work on decision under ambiguity and three-valued logic.[3]

In other words, when I asked us to "take the con out of econometrics" I was only saying the obvious: If the range of inferences that can reasonably be supported by the data we have is too wide to point to one and only one decision, we need to admit that the data leave us confused. Thus my contribution to econometrics has been confusion! You, though, refuse to admit that you are just as confused as I.

Three-valued logic seems especially pertinent when extending the results of experiments into other domains—there is a lot of "we don't know" there. Even in the case of mechanical systems, with statistical properties much better understood than economics/financial human systems, it is not assumed that models will work under extreme conditions. Engineers may design aircraft that according to their computer models can fly, but until real airplanes are actually tested in normal and stressful conditions, the aircraft are not certified to carry passengers. Indeed, Boeing's composite plastic 787 Dreamliner has been suffering production delays because wing damage has shown up "when the stress on the wings was well below the load the wings must bear to be federally certified to carry passengers" (Gates, 2009). Too bad we couldn't have stress-tested those mortgage-backed securities before we started flying around the world with them. Too bad the rating agencies did not use three-valued logic with another bond rating: "AAA-H," meaning hypothetically AAA according to a model but not yet certified as recession-proof.

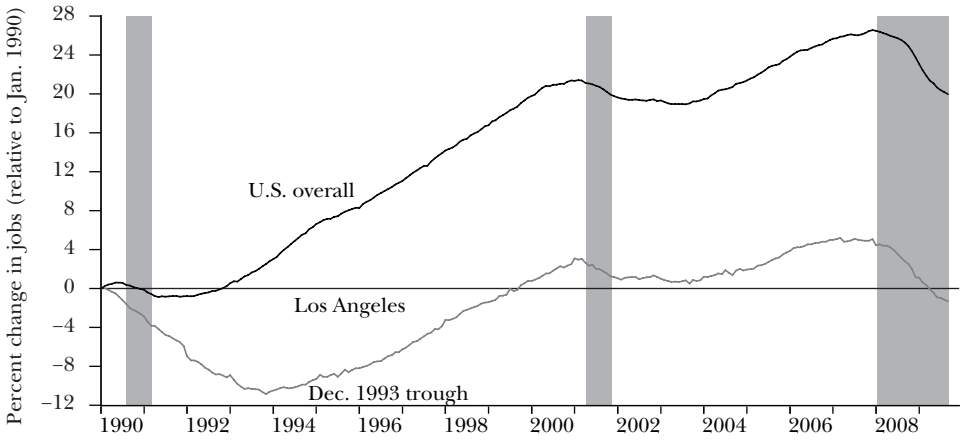
There is of course more than one reason why mortgage-backed securities didn't fly when the weather got rough, but it will suit my purposes here if I make the argument that it was an inappropriate extrapolation of data from one accidental experiment to a different setting that is at the root of the problem. To make this argument, it is enough to look at the data in Los Angeles. Don't expect a full econometric housing model. It's only a provocative illustration.

Figure 1 contrasts nonfarm payrolls in Los Angeles and in the United States overall during the last three recessions. In 2001 and in 2008, the L.A. job market

[2] Please don't think I am assigning a uniform distribution over this interval, since if that were the case, the probability would be precisely equal to the mean: 0.10.

[3] Bewley (1986), Klibanoff, Marinacci, and Mukerji (2005), and Hanany and Klibanoff (2009) and references therein are exceptions.

Figure 1
Payroll Jobs in Los Angeles and U.S. Overall
(*recessions shaded*)



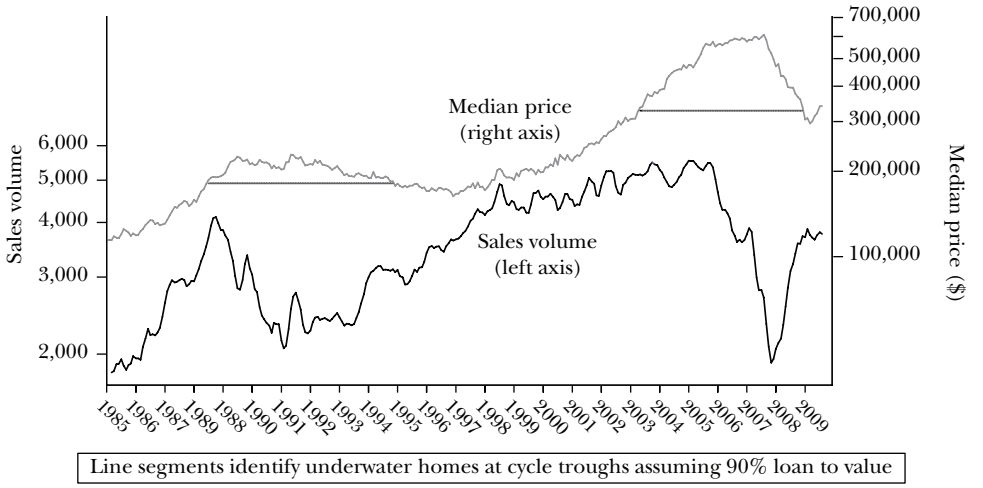
Note: The figure shows the change in number of employees on nonfarm payrolls, in the Los Angeles area and in the U.S. overall, relative to January 1990, seasonally adjusted. The Los Angeles area is the Los Angeles–Long Beach–Santa Ana, California, Metropolitan Statistical Area.

declined in parallel with the U.S. decline, but the recession of 1990–91 was especially severe in Los Angeles. U.S. payroll jobs dropped by 1.5 percent during this recession, while jobs in the Los Angeles metropolitan statistical area declined by 11 percent and did not return to their 1990 levels until 1999. This was a natural experiment known as the end of the Cold War, with Los Angeles treated and with, for example, San Francisco in the control group. (The 2001 recession had the treatment reversed, with San Francisco treated to a tech bust but Los Angeles in the control group.)

Thus, I suggest, the L.A. data in the early 1990s is a test case of the effect of a severe recession on mortgage defaults. Figure 2 illustrates the number of homes sold and the median prices in Los Angeles from January 1985 to August 2009. The horizontal line segments indicate the period over which all homes purchased with a loan to value ratio of 90 percent (indicating a 10 percent downpayment) and interest-only payments will be underwater at the next trough. If all these homes were returned to the banks under the worst case scenario when the price hits bottom, the bank losses are 90 percent of the gap between that line segment and the price at origination. Clearly the underwater problem is much more intense in the latest downturn than it was in the 1990s.

In the first episode, volume peaked much before prices, falling by 50 percent in 28 months. Though volume was crashing, the median price continued to increase, peaking in May 1991, 101 percent above its value at the start of these data in January 1985. Then commenced a slow price decline with a trough in

Figure 2
Los Angeles Housing Market: Median Home Price and Sales Volume



Source: California Association of Realtors.
Note: The median price and sales volume are a 3-month moving average, seasonally adjusted. The line segments indentify homes that will be underwater at the coming cycle trough assuming 90 percent loan to value at origination and interest only payments. (Denoting the price at time t by p_t and the price at the trough by $p_{\min}$, with 10 percent down, the loan balance is 0.9 p_t, and the home is underwater at the trough if 0.9 $p_t > p_{\min}$ or if $p_t > p_{\min}/0.9$.)

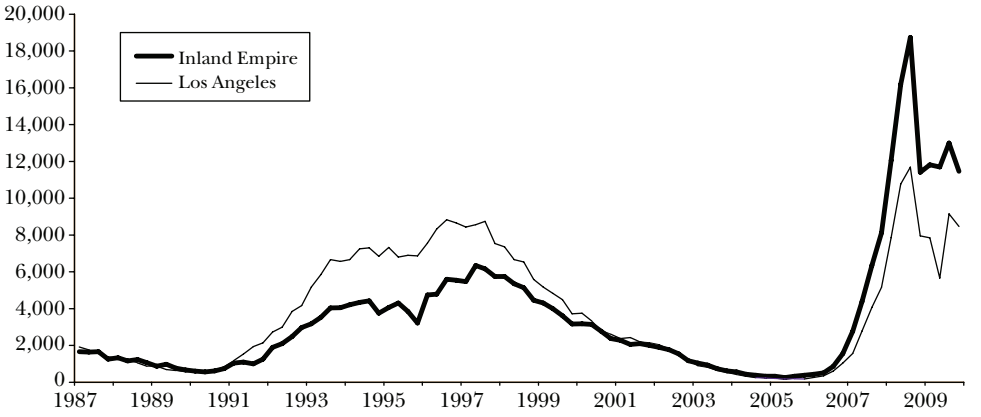
December 1996, with the median price 29 percent below its previous peak, a decline of 5 percent per year.

I have been fond of summarizing these data by saying that for homes, it's a volume cycle, not a price cycle. This very slow price-discovery occurs because people celebrate investment gains, but deny losses. Owner-occupants of homes can likewise hold onto long-ago valuations and insist on prices that the market cannot support. Because of that denial, there are many fewer transactions, and the transactions that do occur tend to be at the seller's prices, not equilibrium market prices.

The slow price discovery acts like a time-out, allowing the fundamentals to catch up to valuations and keeping foreclosure rates at minimal levels.[4] In the early phase of the current housing correction, history seemed to be repeating itself, since volume was dropping rapidly even as prices continued to rise. But then began a rapid 51 percent drop in home prices between August 2007 to March 2009, creating a huge amount of underwater valuation.

[4] The first underwater problem illustrated in Figure 2 could be almost completely remedied by a switch from 90 to 80 percent loan-to-value ratios starting when volumes began to drop, because most of the underwater loans came after that break in the market.

Figure 3
Southern California Foreclosures



Source: MDA DataQuick.
Note: The concentration of foreclosures in time and space in Southern California in the latest housing correction is illustrated in Figure 3, which displays quarterly foreclosures since 1987 for both Los Angeles County and "The Inland Empire," composed of the two "peripheral" counties east of Los Angeles County (Riverside and San Bernadino), where subprime lending was very prevalent. In both Los Angeles and the Inland Empire in the 1990s, the rise in foreclosures trailed the price movement by several years. In contrast, the time-concentrated spike in foreclosures in 2007 occurred at the very start of the price erosion and has been much more extreme in the Inland Empire than in Los Angeles.

Why did the L.A. 1990s data mislead with regard to the current housing correction? One possible answer is untested social effects and unmeasured subject effects—two very important interactive confounders. Innovations in mortgage origination in 2003–2005 extended the home-ownership peripheries of our cities both in terms of income and location. When the subprime mortgage window shut down, the demand for owner-occupied homes at the extended peripheries was eliminated virtually overnight. Since these properties were financed with loans that could only work if the houses paid for themselves via appreciation, the banks became the new owners. Accounting rules do not allow banks to deny losses the way owner-occupants do, and banks immediately dumped the foreclosed homes onto the market at the worst time, in the worst way, with broken windows and burned-up lawns, concentrated in time and space, causing very rapid (or even exaggerated) price discovery in the affected peripheries. In contrast, foreclosures in the 1990s, illustrated in Figure 3 for the case of Southern California, were delayed, and were dispersed in time and space.

We all need to learn both narrower and broader lessons here. I expected price discovery in housing markets to be slow, as it had been after the bursting of previous bubbles, and I was completely wrong. I will be more careful about interactive social confounders in the future.

White-Washing

It should not be a surprise at this point in this essay that I part ways with Angrist and Pischke in their apparent endorsement of White's (1980) paper on how to calculate robust standard errors. Angrist and Pischke write: "Robust standard errors, automated clustering, and larger samples have also taken the steam out of issues like heteroskedasticity and serial correlation. A legacy of White's (1980) paper on robust standard errors, one of the most highly cited from the period, is the near-death of generalized least squares in cross-sectional applied work."

An earlier generation of econometricians corrected the heteroskedasticity problems with weighted least squares using weights suggested by an explicit heteroskedasticity model. These earlier econometricians understood that reweighting the observations can have dramatic effects on the actual estimates, but they treated the effect on the standard errors as a secondary matter. A "robust standard" error completely turns this around, leaving the estimates the same but changing the size of the confidence interval. Why should one worry about the length of the confidence interval, but not the location? This mistaken advice relies on asymptotic properties of estimators.[5] I call it "White-washing." Best to remember that no matter how far we travel, we remain always in the Land of the Finite Sample, infinitely far from Asymptopia. Rather than mathematical musings about life in Asymptopia, we should be doing the hard work of modeling the heteroskedasticity and the time dependence to determine if sensible reweighting of the observations materially changes the locations of the estimates of interest as well as the widths of the confidence intervals.

Estimation with instrumental variables is another case of inappropriate reliance on asymptotic properties. In finite samples, these estimators can seriously distort the evidence for the same reason that the ratio of two sample means, $\overline{y}/\overline{x}$, is a poor summary of the data evidence about the ratio of the means, $\mathrm{E}(y)/\mathrm{E}(x)$, when the finite sample leaves a "dividing-by-zero-problem" because the denominator $\overline{x}$ is not statistically far from zero. This problem is greatly amplified with multiple weak instruments, a situation that is quite common. It is actually quite feasible from a Bayesian perspective to program an alert into our instrumental variables estimation together with remedies, though this depends on Assumptions. In other words, you have to do some hard thinking to use instrumental variables methods in finite samples.

As the sample size grows, concern should shift from small-sample bias to asymptotic bias caused by the failure of the assumptions needed to make instrumental variables work. Since it is the unnamed, unobserved variables that are the source of the problem, this isn't easy to think about. The percentage bias is small if

[5] The change in length but not location of a confidence interval is appropriate for one specialized covariance structure.

the variables you have forgotten are unimportant compared with the variables that you have remembered. That's easy to determine, right?

A Word on Macroeconomics

Finally, I think that Angrist and Pischke are way too optimistic about the prospects for an experimental approach to macroeconomics. Our understanding of causal effects in macroeconomics is virtually nil, and will remain so. Don't we know that? Though many members of our profession have jumped up to support the $787 billion stimulus program in 2009 as if they knew that was an appropriate response to the Panic of 2008, the intellectual basis for that opinion is very thin, especially if you take a close look at how that stimulus bill was written.

The economists who coined the DSGE acronym combined in three terms the things economists least understand: "dynamic," standing for forward-looking decision making; "stochastic," standing for decisions under uncertainty and ambiguity; and "general equilibrium," standing for the social process that coordinates and influences the actions of all the players. I have tried to make this point in the title of my recent book: *Macroeconomic Patterns and Stories* (Leamer, 2009). That's what we do. We seek patterns and tell stories.

Conclusion

Ignorance is a formidable foe, and to have hope of even modest victories, we economists need to use every resource and every weapon we can muster, including thought experiments (theory), and the analysis of data from nonexperiments, accidental experiments, and designed experiments. We should be celebrating the small genuine victories of the economists who use their tools most effectively, and we should dial back our adoration of those who can carry the biggest and brightest and least-understood weapons. We would benefit from some serious humility, and from burning our "Mission Accomplished" banners. It's never gonna happen.

Part of the problem is that we data analysts want it all automated. We want an answer at the push of a button on a keyboard. We know intellectually that thoughtless choice of an instrument can be a severe problem and that summarizing the data with the "consistent" instrumental variables estimate when the instruments are weak is an equally large error.[6] The substantial literature on estimation with weak instruments has not yet produced a serious practical competitor to the usual

[6] Bayesians have a straightforward solution in theory to this problem: describe the marginal likelihood function, marginalized with respect to all the parameters except the coefficient being estimated. If the instruments are strong, this marginal likelihood will have its mode near the instrumental variables estimate. If the instruments are weak, the central tendency of the marginal likelihood will lie elsewhere, sometimes near the ordinary least-squares estimate.

instrumental variables estimator. Our keyboards now come with a highly seductive button for instrumental variables estimates. To decide how best to adjust the instrumental variables estimates for small-sample distortions requires some hard thought. To decide how much asymptotic bias afflicts our so-called consistent estimates requires some very hard thought and dozens of alternative buttons. Faced with the choice between thinking long and hard versus pushing the instrumental variables button, the single button is winning by a very large margin.

Let's not add a "randomization" button to our intellectual keyboards, to be pushed without hard reflection and thought.

■ *Comments from Angus Deaton, James Heckman, Guido Imbens, and the editors and referees for the JEP (Timothy Taylor, James Hines, David Autor, and Chad Jones) are gratefully acknowledged, but all errors of fact and opinion, of course, remain my own.*

References

Bewley, Truman F. 1986. "Knightian Decision Theory." Parts 1 and 2. Cowles Foundation Discussion Papers No. 807 and 835.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review*, 43(2): 245–57.

Chamberlain, Gary, and Guido Imbens. 1996. "Hierarchical Bayes Models with Many Instrumental Variables." NBER Technical Working Paper 204.

***The Economist*.** 2009. "Cause and Defect." August 13, 2009.

Deaton, Angus. 2008. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." The Keynes Lecture, British Academy, October, 2008.

Fox, Justin. 2009. *The Myth of the Rational Market. A History of Risk, Reward and Delusion on Wall Street.* New York: HarperCollins.

Gates, Dominic. 2009. "Boeing 787 Wing Flaw Extends Inside Plane." *Seattle Times*, July 30. http://seattletimes.nwsource.com/html/boeingaerospace/2009565319_boeing30.html.

Hanany, Eran, and Peter Klibanoff. 2009. "Updating Ambiguity Averse Preferences." *The B.E. Journal of Theoretical Economics*, vol. 9, issue 1 (Advances), article 37.

Heckman, James J. 1992. "Randomization and Social Program Evaluation." In *Evaluating Welfare and Training Programs*, eds. Charles Manski and Irwin Garfinkel, 201–230. Cambridge, MA: Harvard University Press. (Also available as NBER Technical Working Paper 107.)

Imbens, Guido W. 2009. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." NBER Working Paper No. w14896.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–75.

Keynes, John Maynard. 1921. *A Treatise on Probability.* London: Macmillan.

Keynes, John Maynard. 1936. *The General Theory of Employment, Interest and Money.* Macmillan.

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2005. "A Smooth Model of Decision Making Under Ambiguity." *Econometrica*, 73(6): 1849–1892.

Knight, Frank. 1921. *Risk, Uncertainty and Profit.* New York: Houghton Mifflin.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* John Wiley and Sons.

Leamer, Edward E. 1983. "Let's Take the Con

Out of Econometrics." *American Economic Review*, 73(1): 31–43.

Leamer, Edward E. 1985. "Vector Autoregressions for Causal Inference?" In Karl Brunner and Allan H. Meltzer, eds., *Carnegie-Rochester Conference Series on Public Policy*, vol. 22, pp. 255–304.

Leamer, Edward E. 2009. *Macroeconomic Patterns and Stories*. Springer Verlag.

Liu, Ta-Chung. 1960. "Underidentification, Structural Estimation, and Forecasting." *Econometrica*, 28(4): 855–65.

Sala-i-Martin, Xavier. 1997. "I Just Ran Two Million Regressions." *American Economic Review*, 87(2): 178–83.

Sims, Christopher. 1980. "Macroeconomics and Reality." *Econometrica*, 48(1): 1–48.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 48(4): 817–38.

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. G. W. Harrison. 2011. Randomisation and Its Discontents. *Journal of African Economies* **20**:4, 626-652. [CrossRef]
4. G. W. Harrison. 2011. Experimental methods and the welfare evaluation of policy lotteries. *European Review of Agricultural Economics* . [CrossRef]
5. Judea Pearl. 2011. Statistics and Causality: Separated to Reunite-Commentary on Bryan Dowd's "Separated at Birth". *Health Services Research* **46**:2, 421-429. [CrossRef]
6. T. Plumper, V. E. Troeger. 2011. Fixed-Effects Vector Decomposition: Properties, Reliability, and Instruments. *Political Analysis* **19**:2, 147-164. [CrossRef]
7. Henk Folmer, Olof Johansson-Stenman. 2011. Does Environmental Economics Produce Aeroplanes Without Engines? On the Need for an Environmental Social Science. *Environmental and Resource Economics* **48**:3, 337-361. [CrossRef]
8. C. B. Barrett, M. R. Carter. 2010. The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections. *Applied Economic Perspectives and Policy* **32**:4, 515-548. [CrossRef]
9. 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]