Econometric Data Analysis: A Cost-Effective, Largely Successful Methodology

By Lee S. Friedman
Professor of Public Policy
The Richard and Rhoda Goldman School of Public Policy
University of California at Berkeley
2607 Hearst Avenue
Berkeley, CA 94720-7320
Email: lfried@uclink.berkeley.edu
Phone: (510) 642-7513
Fax: (510) 643-9657
Home Page: http://webdisk.berkeley.edu/~lfried

April 2005

This paper is a draft prepared for purposes of discussion at the National Academy of Sciences—National Research Council conference on “Use of Evidence for Social Policy Decision Making” in Irvine, CA (April 1-3, 2005). Please do not cite or quote without permission from the author.

Econometric Data Analysis: A Cost-Effective, Largely Successful Methodology

I. Introduction

This paper is intended to convey the role of econometric data analysis as one method used in the social sciences to provide factual evidence. Econometrics is, generally speaking, the set of statistical procedures used to estimate economic models. The procedures are used to explain and predict the levels of economic variables as well as to test hypotheses about their relationships, and the results are often used as evidence in a wide range of policy settings.[1]

The specific subjects of these studies cover virtually all parts of economic theory: macroeconomic phenomena like understanding the causes and consequences of inflation, employment levels and savings rates; and microeconomic phenomena like estimating demand and supply curves for particular goods and services, and understanding the behavior of consumers and the production decisions of firms. They also include the effects of public policies in all of these areas.

Economists using these methods have not been particularly shy about extending their use to areas that are not always thought of as economic: understanding suicide, marriage and divorce rates, criminal behavior, and the results of childhood education efforts. Economic theory often offers some new insight, for example, that the activity can be thought of as having a price and that its extent will depend upon the price level that people have to pay for it. In crime, for example, the price is the expected punishment. The very first econometric efforts to test the theory are often sharply criticized by those who have studied the areas far longer, but over time the economists become more sophisticated by addressing these criticisms. In the cases mentioned, eventually econometric contributions are acknowledged, even if grudgingly so, by the noneconomist critics.

[1] Many textbooks are designed to teach econometric methodology. Examples include Jack Johnston and John Dinardo, Econometric Methods, 4th ed., McGraw-Hill (Boston, MA: 1998), Fumio Hayashi, Econometrics, Princeton University Press (Princeton, NJ: 2000) , Robert S Pindyck and Daniel L Rubinfeld, Econometric Models and Economic Forecasts, McGraw-Hill (1997), and G.S. Maddala, Introduction to Econometrics, John Wiley & Sons (New York, NY: 2001).

One important aspect of econometric procedures is that they have largely developed on the assumption that the data would be generated from naturally-occurring activity, rather than from a formal experiment in which a specific treatment is to be tested in order to determine its effects. In the experiment, the effects of all factors other than the treatment of interest are intended to be removed by the process of random assignment to an experimental or control group. Econometric procedures, on the other hand, are intended to account for all nonrandom influences by explicitly incorporating them as variables in the econometric model. Economists may be interested in how the price of natural gas influences home energy consumption, but to estimate this in an econometric model means also including variables accounting for the weather, the size of the home, the number of occupants and their employment or school status, and other factors that influence home energy consumption.

How do we understand the power of these procedures, and what are the limits to them? What are the implications of this in terms of standards of evidence in the social sciences? I'd like to offer the bottom line of my discussion in advance. It is as follows. As much as I love experiments, in a world of limited research resources it is critical to understand the high cost-effectiveness of nonexperimental, econometric methods. These methods complement and provide independent checks on experimental findings. It is crucial not only to continue supporting these efforts, but to continue to support the necessary infrastructure for them: the extensive data collection efforts of our censuses and surveys. For studying social policies, the tradeoff between experimental and nonexperimental methods is something like this: a substantial tilt toward increased experimental research will result in better evidence in some dimensions, but far less of it because many fewer studies could be supported. The advantages of an experiment come from greater internal validity (certainty about the treatment effect) but often are offset by greatly reduced external validity and uncertainty about how to replicate the treatment. Because econometric methods have opposite strengths and weaknesses, the combination of the two approaches is preferable to more exclusive reliance on either alone.

Let me quickly add, to be sure that I am not misunderstood, that I love experiments: I was responsible for the economic evaluation of the original Supported Work social experiment, I have analyzed the impact of a criminal justice experiment designed to increase the number of accused defendants who could be safely released while awaiting trial, and currently I am supervising a doctoral thesis analyzing a recent experiment with electricity pricing in California. I also know that there are many laboratory-type experiments of good research value and relatively low cost, and there are surely some field studies for which it would be worthwhile to convert into field experiments. Nevertheless, my concern here is to be sure that this proposition is understood: to maximize the value of social science research with a budget of any fixed size, a substantial portion of the research portfolio must continue to be allocated to nonexperimental research methods like econometrics.

II. The Promise of Econometric Methods

While econometrics can never provide absolute proof that one factor causes another, it may provide good evidence of causality when (1) a statistical relationship is persuasively documented, and (2) plausible theoretical explanations to explain the relationship are consistent with respect to the direction of cause and effect between the two factors. Of course the words “persuasive” and “plausible” are terms of judgment about which, in the end, reasonable people may disagree. Nevertheless, there is a fairly well-established set of hoops that are used to conduct and to evaluate econometric work, and their widespread use by professionals helps to create agreement and to narrow the range of disagreement. I hope that I can convey, in a short and not too technical exposition, the flavor of this process.

Consider the economic proposition that the demand for a commodity will fall as its price rises, other things being equal. Economists think that other things besides a commodity’s own price might affect the demand, like the price of substitute commodities and the general level of income. So the price proposition is tested econometrically using multiple regression analysis to control for the effects of the other factors. No single study is taken as the convincing proof of this fundamental proposition. But because thousands of such analyses have been done independently, on hundreds of commodities in hundreds of different communities—and because these studies uniformly find that demand falls as price rises—economists agree that the basic proposition is correct.

This simple description hides the great complexity of actually conducting a persuasive study. I'd say that the number of serious issues to be resolved in order to do one is somewhere between dozens and a hundred, although I've never actually tried to count them. Roughly speaking, these issues might be thought of as belonging to three categories: (1) matching theory and hypotheses to available data (the specification problem); (2) making statistical inferences from a particular body of data (model estimation and testing); and (3) drawing appropriate conclusions, including predictions and policy implications, from the estimated model.

The process of controlling for other factors is not necessarily statistical. It does involve mathematical modeling or specification: identifying a precise numerical relationship among several factors. Theory often offers guidance about these relationships: whether variables are positively or negatively related, limited possible ranges for the parameter values, that one parameter must be smaller or larger than another. But rarely does it identify precise values. Statistical procedures help us to identify these numerical relationships when they are imperfectly observed (due to other factors, disturbances, that cause random deviations from the relationship). Statistical tests are used to assess the level of confidence in the relationships established under these imperfect conditions. Even when we are confident in a relationship, we may not be confident in our understanding of the underlying causal mechanism that explains it. More than one theory can be consistent with an established relationship. We may then search for new opportunities where the competing theories offer contradictory predictions in order to test further. Absent such new opportunities, we fall back on judging the plausibility of the alternative theories: perhaps one offers a consistent explanation for a wide range of similar situations, while the other is "new" and has not been tested elsewhere. In this situation, one is more likely to favor the established theory over the new one, although the truth is uncertain.

This might be a good time to mention that the standard for relying upon any particular econometric result depends on the purpose of the user. Most of my discussion is about the standards used by professional economists themselves, evaluating for a purpose something like “what the study has contributed to knowledge”. But decision-making users, like those in the public sector who choose and shape public services, have very different standards of usability. In some decision-making cases, perhaps like choosing a medical treatment, it can be helpful to know if one alternative is slightly more likely to be effective than another, even if the difference is not statistically significant at the “usual” levels. In other cases, estimates that achieve standard statistical significance may still be far too imprecise, as when millions of dollars of tax revenue can be affected by a very small change in the exact tax rate used.

I have tried so far to emphasize the role of theoretical guidance, and the consistency of results with it, in my brief description of factors that determine the persuasiveness of econometric work. Confidence in econometric results depends on far more than the reliability of the specific data and appropriateness of statistical inference methods used to analyze it in any one study. It depends heavily on understanding of and confidence in the underlying theory that has motivated the study. To a large extent, the successes of econometrics reflect the successes of economic theory.

The results of econometric work are used routinely for decision-making in both the private and public sectors. When a large corporation faces a major investment decision like whether or not to build an expensive new plant to expand its capacity, it often uses an econometric model to predict the state of the economy, the expected corporate sales and the likely profitability of the plant. When federal regulators try to assess whether a firm has exercised market power to illegally manipulate prices as in the California electricity crisis, econometric work is used to distinguish whether or not the observed prices can be explained by normal competition or not. When damages due to workplace injuries are to be awarded in court cases, econometric models are often used to establish their magnitudes. Econometric studies are commissioned by the Environmental Protection Agency to estimate the likely cost of a tradable emission permit in the future, based on the economic theory that prices will reflect marginal costs. When nonprofit hospitals propose to merge, econometric studies are undertaken to estimate the likely effect on hospital charges. When changes in the tax code are considered like those reflecting the history of the Earned Income Tax Credit, econometric studies are used to assess the likely changes in work effort of those affected. If the acid test of the value of something is the extent to which people voluntarily buy it, then econometric methodology clearly has high value. The list of applications is essentially endless, and the extensive use underscores the need to continue to improve and advance the state of the art.

III. A Quick Tour of Econometric Issues

It is, of course, impossible to give in a short paper a comprehensive overview of specific econometric issues that must be confronted in the course of an application. I have selected a very small number, in the hope that they will convey the flavor of the task.

A. Matching theory and hypotheses to available data (the specification problem)

Specifying a functional form. Economic theory often suggests the variables that should be included in a theorized relationship, but typically stops somewhere short of specifying the precise mathematical way that the variables are related.

For example, a demand function suggests that the consumption amount Q of a normal good will increase with income Y and decrease with price P, but not necessarily the specific form. Two common functional forms that have this property are linear and log-linear, although there are of course others. Assume that we have observations on the variables, and that there are other minor factors that do not intrinsically concern us but cause small random deviations u from the theoretical relationship. Then we could represent the two common forms:

$$Q = a + bP + cY \quad + u \qquad\qquad (b < 0, c > 0)$$

$$\ln Q = a + b \ln P + c \ln Y + u \qquad (b < 0, c > 0)$$

As long as the disturbances can be assumed to be independently drawn and normally distributed, the above equation parameters a, b, and c could be estimated by ordinary least squares regression. Older studies used to simply assume a specific functional form (often like one of the two above). However, modern practice is to specify a more general functional form that includes the older ones as special cases, and let the data determine the specific form. One method of doing this is to use the Box-Cox transform, which identifies by the maximum likelihood method a specific parameter λ from the range 0 to 1, where 0 is the linear form and 1 is the log-linear form. One can also test a result like $\lambda = .8$ to see if it is significantly different from 1. The point is that a study that is sensitive to this choice of functional form issue is preferred to one that is insensitive to it.

Omitted Variables. To some extent, this is an available criticism of almost any econometric study because it so easy to think of something else that would have been nice to include. An example of quite constructive criticism, however, comes from the education area where early econometric efforts to explain a child’s educational progress focused on school resources only: spending per pupil, class size, teacher quality, etc. Over time researchers learned that important omitted factors included the nature of other students in the class, the student’s family background, and aspects of the student’s home neighborhood.

We might consider as a special case of this category the measurement problem: is the included variable selected to represent a particular influence actually representing that influence? If not, then the true variable is still omitted. What variables, for example, measure the kind of teacher quality relevant to the learning of children? Highest degree? Years of experience? Quality of undergraduate training?

If a relevant variable has been excluded from the analysis, it can bias the estimated coefficients of the included variables. The direction of the bias is given by the sign of the true coefficient on the excluded variable multiplied by the sign of the correlation between the included and excluded variable. There is no bias if the excluded variable is either uncorrelated with the included ones or if its true coefficient is zero. There isn't too much that one can do about omitted variables, in the sense that they are usually omitted because the appropriate observations of them are not available for the sample. However, good practice is to do the following: (1) if crucial variables are known to be missing, do not do the study on that dataset; (2) offer a good discussion of possible omitted variables and the bias that they might cause; and (3) most creatively, use proxy variables that are available to take the place of the variable that would otherwise be omitted. An example of the latter is that sometimes an individual's wealth is more relevant than the current income level for certain purchases, but there are rarely good measures of this wealth. However, sometimes good proxies for wealth are available: the square footage of the home, or income data combined with demographic data like age and education.

Structural homogeneity of the sample. Economists often test the theory of individual behavior using observations on groups of individuals. For example, the economic theory of crime and deterrence is a theory that asserts individual choices will depend, other things equal, on the level of punishment. However, the data available to test this theory is usually based on geographic units like cities, counties, states, or the country as a whole (within which crime rates, arrest rates, etc. are available); the data may or may not involve a time series. The use of these aggregated observations can cause serious bias in the parameter estimates if individuals in one region or time period behave differently than they do in another region or time period. This is a serious problem because the parameters can be biased in either direction, depending on how the true differences among regions are distributed. There will be no bias if individuals in the sample are homogeneous across regions and time. One method of testing for structural homogeneity is to use Chow tests to see if any of the estimated model parameters are significantly different over regions or time, and if so, corrective procedures may require additional dummy variables or separate estimating equations. An econometric study that fully tests for structural homogeneity will be preferable to one that either ignores it or considers as possible controls dummy variables interacted only with the constant term.

B. Making statistical inferences from a particular body of data (model estimation and testing)

Once one has settled on the data and the model, numerous problems may remain before appropriate statistical inferences can be drawn. I mention very briefly three. They all have in common violations of the usual assumption for regression estimation that the error or disturbance terms will be independently and normally distributed.

Simultaneous equations bias. In economics, many observations of market price and quantity outcomes are thought to be jointly determined by demand and supply curves. A very simple representation of them is as follows:

Demand $Q = a + bP + cY + u_1$ $(b < 0, c > 0)$
Supply $Q = d + eP + fZ + u_2$ $(e > 0, f > 0)$

If observed Q and P are determined jointly by both equations, then how does one get an estimate of each?

If one runs ordinary least squares on the equations separately, then the variable P will not be independently distributed from u_1 (because P is not really exogenous; it is jointly determined by the two equations). The estimated parameter b will be biased, and could be anywhere between b and e depending on the relative sizes of the variance of the errors terms u_1 and u_2. In other words, one doesn't know if one has estimated the demand parameter, the supply parameter, or some weighted average of the two. The estimate is likely useless. This problem is very closely related to the identification problem, which is more generally how to identify the individual coefficients in the equations of a simultaneous equation model.

There are a variety of procedures that can be used to solve these problems. Common methods include the use of instrumental variables, indirect least squares, two-stage least squares and others. Again, good econometric work will consider carefully the nature of any simultaneous equations bias, and will undertake corrective procedures appropriate for the specific case.

Heteroskedasticity. The standard assumptions about the error or disturbance term are not only that they are independently and normally distributed, but with constant mean and variance. The assumption of constant variance is violated when the error terms are correlated with one or more of the independent variables, e.g. the size of the errors increase with the income level in estimating a demand equation. This problem is somewhat less serious than the others mentioned in that it does not cause bias in the estimates. However, it does invalidate the usual tests of significance because it biases the variance estimates. Heteroskedasticity may be diagnosed by a variety of tests such as those suggested by Ramsey, White and Goldfeld-Quandt, and corrective procedures may involve using weighted least squares or maximum likelihood methods.

Serial correlation. In many time series studies, the assumption of independent errors is violated by the presence in the model of a lagged dependent variable or an expectations variable that depends upon prior history. In such cases, the error term in any one period is correlated with the error terms in the immediately preceding periods. This results in inefficient and in most cases biased estimates. The presence of serial correlation can be detected by tests like the Durbin-Watson statistic, or Durbin's *h*-test, or tests based on the Lagrange Multiplier principle. Corrective procedures, if serial correlation is found, may involve transforming the data based on estimating the degree of first-order autocorrelation, or estimating the equations by using the first-differences of the sequential observations rather than their absolute levels.

C. Drawing appropriate conclusions, including predictions and policy implications, from the statistical inferences.

Consistency with theory and prior estimates. It is normal, once the models are estimated, to then report the consistency or inconsistency of the results with the underlying theory. For example, do the estimated parameters have the expected signs and are they statistically significant? How do the estimates compare with previous estimates reported in the literature, and what might explain any differences (many studies are undertaken because new and better data have become available, compared to older efforts)?

Competing hypotheses. In addition to the general checks on model consistency with theory and previous estimates, there are often other specific reasons why the model has been estimated. In some cases, the motivation behind the study is to test two alternative theories against each other to see which is more accurate. In macroeconomics, there may be Keynesian versus monetarist theories. In microeconomics, there is much current attention to what is now termed behavioral economics: the applicability of various models of limited or bounded rationality to economic decision-making. Growing attention will be paid to whether models based on conventional or behavioral theories explain actual decisions better.

When alternative theories are to be tested against one another using the same data set, the nature of the appropriate test depends on the specific source of differences between the alternative models. I'll mention briefly the J-test, which is appropriate when used with two different, non-nested theories (the variables of one are not a subset of the other) to explain the level of the same dependent variable. It evaluates which model is better by asking if one model adds any significant new explanatory power to the other (and vice-versa). It essentially adds to the model being tested an additional right-hand-side variable equal to the predicted level of the dependent variable by the other model. If this new variable is significant based upon its t-statistic, then one rejects the model being tested in favor of the alternative. It is possible, however, for the results to be ambiguous: both models could be rejected, and both can be maintained.

Another procedure that may be used to evaluate competing hypotheses is to see which one predicts better on a sample unused during estimation, sometimes called crossvalidation. This sample cannot be a strictly random subset of the original data, because if it were the results would be essentially identical to those using the estimating sample. I had a very interesting case of this, in testing a behavioral economic model (BR) against a conventional one (SUM) in the context of residential energy consumption.[2] Both models predicted identical consumption at relatively low levels (prices at lifeline levels), but the BR hypothesis predicted consumption greater than SUM at levels above lifeline quantities. Therefore, I reserved a random portion of observations for prediction purposes drawn only from those in the upper 80 percent by square footage (i.e. tilted toward households likely to be higher energy consumers). I estimated the two models from observations randomly drawn from the full sample. Then I compared the predictions from the two models by calculating the root mean square error of each applied to the prediction sample; BR, with the lower root mean square error, was the winner.

Policy Uses. There are many uses covering many quite different decision-making circumstances, and it is difficult to describe general rules for high standards in such diverse circumstances. Competition is one good thought to mention. When time permits, a number of independent studies might be commissioned and then the results of each scrutinized relative to one another. This is the case in many courtroom and regulatory proceedings, in which different sides or interest groups will present differing estimates of the consequences of some policy action like the estimated effect on prices if a merger of two entities is permitted. Even if a single econometric study is offered as evidence for a particular policy position, it is usually desirable to have it assessed by some other independent expert. The greater the stakes, the more effort it is worth to narrow the range of uncertainty about econometric estimates. However, I am also mindful of urgent situations in which a decision must be made quickly (like my choice of medical treatment example earlier), where perhaps standards lower than the usual statistical ones might be of high value.

[2] See Lee S. Friedman, “Bounded Rationality versus Standard Utility-Maximization: A Test of Energy Price Responsiveness,” in R. Gowda and J. Fox, eds., Judgments, Decisions, and Public Policy, Cambridge University Press (New York, NY: 2002), pp. 138-173.

From another angle, how might authors of econometric studies better guide any appropriate policy uses of them? It is generally routine to discuss the generalizability of the findings, and that is in itself very helpful. More explicit attention to the questions of potential policy users could be very helpful. One might work very hard and very well to produce an unbiased estimate of the extent to which capital punishment deters murder (we still do not know that it does); what matters from a policy perspective however, is the extent to which it deters murder over and above life imprisonment. In a study of the production levels of pretrial service agencies, I tried to discuss the results from the perspective of a manager of one of these agencies: what variables represented matters of judgment about when and where the service should be produced, and which were crucial to understanding the overall effectiveness with which the service was delivered.[3]

Let me turn now to the final section of my discussion, concerning experimental versus econometric methods.

III. Experimental Versus Econometric Approaches[4]

I mentioned earlier alternative behavioral economic models that emphasize the difficulty or the impossibility of obtaining and processing the information required to maximize utility. One attribute of them relevant to note here is that the existing tests of behavioral hypotheses almost always rely on data that is not normally observable in naturally-occurring market settings. There are many studies based upon laboratory experiments that show subjects behaving inconsistently with utility-maximization, but that do not test any other specific theory.[5] However, SUM models continue to guide most applied econometric research.

[3] See Lee S. Friedman, “Public Sector Innovations and their Diffusion: Economic Tools and Managerial Tasks,” in A. Altshuler and R. Behn, eds., Innovation in American Government: Challenges, Opportunities, and Dilemmas, The Brookings Institution (Washington, DC: 1997), pp. 332-359.

[4] An earlier version of these thoughts is presented in Friedman (2002), op. cit.

[5] See V. L. Smith, “Theory, Experiment and Economics,” Journal of Economic Perspectives, *3*, 1989, pp. 151-169 for an introduction and summary of this literature. The same generalization can be applied to the public goods literature studying actual preference revelation more "honest" than expected by conventional theory; see, for example, G. Marwell and R.E. Ames, “Economists Free Ride, Does Anyone Else? Experiments on the Provision of Public Goods,” Journal of Public Economics, *15*, 1981, pp. 295-310.

A. The Experiment has many Advantages

There is much to be said in favor of the formal experimental methodology. One has the ability to plan treatments and observe the responses to them that may be difficult or impossible to observe in natural market settings. The evidence concerning preference reversals is a good example of this.

But what about situations in which one can observe important evidence in the actual market? This is particularly relevant for public policy research, because the effect of actual policies in the marketplace is paramount. Another important advantage of the experiment is the ability to design highly precise treatment effects. That is, the variation in decisions across groups is fully attributable to the designed treatment differences among the groups, save for some small statistical noise. By contrast, the use of natural (non-experimental) decisions in the marketplace requires the analyst to account for all of the factors that may explain systematic differences in choices among individuals. Such studies, as we have seen, are always subject to the criticism that important factors besides the "treatment" of interest have not been sufficiently controlled, possibly leading to biased estimates of the "treatment" effect.

For example, suppose we wish to know how consumers respond to a price increase for a given product. The experimentalist will assign people randomly to a treatment group that will face the higher price, and a control group that will not. The experimentalist will conclude, subject to normal statistical inference, that the price increase causes the difference in average consumption between the two groups.

The analyst who uses non-experimental market data, however, will first have to make sure that the data includes both the consumption of individuals who have and who have not experienced the price increase. Typically this will involve time-series data from one geographical area, or geographic cross-sectional data within a given time period, or a combination. If, say, the price difference occurs across regions, then the analyst must make sure to control for non-price regional factors that may cause differences in consumption (e.g. if studying home energy consumption, control for climate, wealth, and residence size differences). Similarly, there may be non-price factors that cause changes in consumption over time (e.g. weather, changes in household size). The list of non-price factors may be large, and the ability to get data that measure each of these differences accurately may be limited. And of course the treatment studied—the size of the price increase—is limited to what has actually happened, rather than chosen by experimental design.

B. The Experiment has Important Weaknesses

Why, given the complication and imperfection of studying non-experimental market decisions, would the researcher ever prefer to study them? There are several important reasons. If the alternative is a non-experimental survey, then it is well-recognized that survey responses are not always reliable indicators of how people behave when making actual decisions. This uncertainty makes it valuable to know if survey-based findings are consistent with what can be observed in actual market settings. Indeed, experimentalists value this as well, because there are varying degrees of undesirable artificiality in experiments.

Undesirable artificiality (or weak external validity). To clarify this, let us note the important distinction between a "laboratory" experiment, which is the predominant mode for behavioral economic research, and a "social" experiment. The "laboratory" experiment in economics typically uses as subjects university students who differ from the actual decision-makers in the “real” setting. Almost by definition, the laboratory experiment almost always takes place in an environment or setting that is quite unlike the naturally-occurring market or policy environment.

The energy consumption decisions that I studied would be difficult to simulate in a laboratory experiment. The decisions involve consumer responses to energy price schedules, but it is more than that. A key part of the actual decision environment is the long period of time, perhaps several years, during which the consumer forms a routine by making a series of (daily) decisions with infrequent and limited feedback about the consequences (the monthly bill). The lag between the decisions and feedback is (perhaps) long enough to forget important circumstances that framed the original decisions, and long enough so that the circumstances for the next series of decisions may have changed substantially from the prior series.

The "social" experiment, by contrast, takes place in its natural setting with subjects who would normally be found in this setting: e.g. actual residential energy consumers who get actual bills, or actual low-income families who would qualify for a welfare program being studied. In principle, the earlier criticisms of the non-experimental market study can be avoided by the "social" experiment.

High Costs. But few "social" experiments are conducted because they are monetarily expensive, difficult to arrange, sometimes raise difficult ethical issues, and time consuming. If these costs were no object, almost all experimentalists would prefer to conduct a "social" rather than a "laboratory" experiment to understand choice observable in the marketplace. However, given the limited amount of resources available for research, it is only possible to conduct "social" experiments in rare circumstances. These usually involve high public policy stakes that make the cost of the social experiment seem small in comparison. For my study of residential home energy consumption, it was not of high enough social importance to seek a large, long and costly social experiment. While I remain open to the cleverness of laboratory experiment designers to test the models in the home energy context, I have already described why I think this might not be possible. I believe that it was well worthwhile to study the actual decisions with non-experimental methods. This was reinforced by the unusual comprehensiveness of the data available, as well as the direct policy relevance of evidence on the effects of rates set by regulatory commissions.

Even when social experiments are conducted, important elements of artificiality often remain. For example, an experimental price increase is often regarded by participants as more temporary than a naturally-occurring market price increase. This can affect the investment choices of subjects. In the negative income tax social experiments, subjects whose real wage rate increased through lower taxes might have invested more in income-producing education if the increase was permanent. In our energy price increase example, fewer new energy-efficient furnaces will be bought under a temporary social experiment than under an equal-size actual market price increase.

Selection bias (creaming). Another source of artificiality in the social experiment is that the experimental population, while consisting of real market participants, may not be representative of the broader population to which the treatment may be administered. A rigorous social experiment that recruits volunteer participants who are randomly assigned to experimental and control groups may not tell us much about the effects of the same treatment when applied to those who did not volunteer. Similarly, the environment in which the social experiment takes place can have important effects: the response of participants in a suburb might be quite different from the response of an otherwise-identical group located in a large city.

Not understanding the treatment (replication difficulty). Many social experiments leave as a mystery just what it is about the treatment that caused the result, whereas unraveling this mystery can have enormously important economic consequences. The pretrial service agencies that I studied and mentioned earlier all were stimulated by a successful social experiment in New York City known as the Manhattan Bail Project. Run by the Vera Institute of Justice, it was shown that the Project greatly reduced the number of criminal defendants detained while awaiting trial, and ensured that they appeared as required. But just what did they do to achieve this result, and what would someone else have to do to replicate it? Were the defendants released because the Project really did identify those who could be trusted to appear, or were they released because a well-dressed Vera attorney impressed the judge? Did the specific point system used by the Project matter? The efforts to verify a defendant’s community ties? The follow-up system used by Vera to remind the defendants when and where to appear? These questions were not addressed, and when the number of agencies providing this service spread across the country, many of them operated with very little effectiveness because of differences from the initial experiment. I think my work demonstrated that plausible answers to most of these questions could be provided by an econometric study using combined cross-section and time-series data.

C. Econometric and Experimental Methodologies Complement One Another

All of these sources of artificiality and uncertainty can, in principle, be removed (or at least reduced) by more complex, larger, more inclusive, and longer-duration social experiments. But then we run into the cost issue again. Once one recognizes that the costs prohibit us from routinely doing the "ideal" social experiment, we have several different ways of lowering costs. Some of them retain the experimental design, but move from larger to smaller social experiments and then down to laboratory experiments, with each step increasing the artificiality and reducing the generalizability of the observed decision-making. Alternatively, we can move away from the experiment but retain some of the comprehensiveness of time, place and population studied by using the non-experimental research design.

Imagine research efforts of both types that have equal (and relatively low) costs. It is not at all clear which is preferable. One must judge the extent of artificiality in the experiment against the quality and comprehensiveness of the data available for the non-experimental design. Because each method has different strengths and weaknesses, it is valuable when possible to know if the findings are consistent across them. That is, I suspect that in most cases, one learns more from doing both kinds of research than from concentrating on only one of the two approaches.

IV. Conclusion

I have tried to convey the flavor of econometric methodology, and of the care that is required to execute such a study to high standards. Perhaps the most important aspect to understand is how closely linked to economic theory good econometrics must be. It is typically the theory that guides the preliminary specification of the model, provides the motivation for what is to be tested, and helps to evaluate the results. Once one has a theoretical model and a dataset, there is a tremendous amount of further specification and empirical diagnosis that must be done in order to end up with good estimates. Then the use of the model for hypothesis testing or prediction must be carefully done as well. Even a very good econometric model cannot remove uncertainty about its conclusions. That is why many policy uses of econometric work involve the comparison of several independent studies or at least scrutiny by an independent expert. Some sensitivity to the potential policy uses of an econometric study can produce better discussion of the study's implications and generalizability.

Some researchers think that experimental methods should always be preferred to nonexperimental methods. This is not true in any world in which the budget for research has some limit to it. The strength of the experiment is its internal validity and the ability to design treatments precisely, but limited budgets can put severe limits on the external validity or the ability to generalize beyond the experimental setting. The laboratory experiment is quite unlike the real world of economic decision-making, and even very expensive social experiments have important elements of artificiality and leave as a mystery just what aspects of the treatment might be important for replication. Econometric studies are in widespread use both in the marketplace and by government because they are a generally cost-effective research strategy and because they offer complementary strengths and weaknesses relative to the experiment. We should continue to strive to improve them and the data sources that are central to their use.