

Are CGE Models Reliable?

Ed Leamer

October 13, 2015

Given my absence from one day of deliberations on October 21, and my attendance remotely on the other day, I decided to offer some thoughts in writing about the role of theory and evidence in making public policy, contrasting the CGE (computable general equilibrium) calibration approach with the econometric (study of correlations) approach. These comments are fundamentally about reliability: What methods of creating knowledge support the most reliable public policy advice? What is the best way to communicate to policy makers when the reliability is a special concern and when it is not?

Reliability Concerns In the EPA Charge Documents Are not Prominent Enough

Concerns about reliability seem to be in the background of the EPA documents provided for our deliberations, but should be in the foreground, I suggest. The EPA documents are focused on questions like: What kind of model allows us to include feature X? What kind of model is best suited to answering question Y? These are critical questions, but not the only questions. For example, in studying the impact of NAFTA on US wages, it might be wise to have one model for competition between the US and Mexico and another model for competition between the US and Canada. But deciding which feature needs to be included in each model is not the end of the enterprise. The next step is to determine if an estimated model can be relied on to make policy decisions. A features-driven discussion tends to lead to large Rube Goldberg models with thousands of moving parts, but concern about reliability tends to favor smaller models which are more understandable and less likely to behave in surprising and inappropriate ways, for example, when dynamic properties or nonlinearities that work fine in the historical domain of the model behave wildly in potential future domains.

DSGE models are inherently unreliable

The “Memo on Using Other (Non-CGE) Economy-Wide Models to Estimate Social Cost of Air Regulation” (September 2015) relegates reliability concerns to a final paragraph preceding the Concluding Remarks (p. 10) in which the poor forecasting record of DSGE models is observed. A startling attack on DSGE models is reported there:

“Caballero (2010) echoes this sentiment: “macroeconomics—by which I mainly mean the so-called dynamic stochastic general equilibrium approach—has become so mesmerized with its own internal logic that it has begun to confuse the precision it has achieved about its own world with the precision that it has about the real one.”

I would have thought that this would lead to considerable concern about using DSGE models in the environmental policy-making setting, but the EPA document instead adopts a lawyerly approach (innocent until proven guilty):

“That said, it remains an open question – to our knowledge, undiscussed in the literature – to what extent this criticism applies when DSGE models are used to assess the potential welfare effects of a policy ex-ante rather than for forecasting purposes. When conducting benefit-cost analysis of air regulations, the EPA focuses on estimating changes in welfare relative to a baseline rather than correcting predicting any particular future.”

There are important differences between problems of forecasting and control, but the Caballero comment is not about forecasting, it’s about a model-building culture which allows any goofy model to be created provided it satisfies the requirements of the DSGE acronym. It’s a feature-driven culture, with the complete insistence on the set of features that allow the DSGE label. In that literature, it is hard to find serious discussions of reliability, for forecasting or for control. There often isn’t even any concern expressed about “suitability.” Are “dynamic” and “stochastic” and “general equilibrium” features suited to the enterprise of determining the benefits and costs of air quality regulations?

As for reliability, it is my view that DSGE models take the bad things about CGE models to the next level, capturing in the acronym the three things that economists most poorly understand: Dynamic, how the human system thinks about the future; Stochastic, how the human system deals with statistical uncertainty (assuming away model ambiguity); and General Equilibrium, how all the components of the human system interact. Best, I suggest, would be first to have a conversation about how the effects of environmental policy might depend on forward-looking behaviors, choices under statistical uncertainty and systemic effects. Then the model needs to be built to encompass the specific problems identified or else the modelling energy should focus elsewhere if these are not the critical issues. The goal is not truth. The goal is reliability. Like a road map. A road map is a model of the surface of the earth explicitly designed for efficient decisions of one particular type but it will work terribly for others. We need to know when it works and when it does not. For example, the maps that display the Los Angeles freeways in bright colors as if to recommend their use are discovered to be quite misleading during the many hours of rush hour.

By the way, the only way to tell if a map is useful or not is to use it. Falsification is irrelevant since a map derives its usefulness from not being a perfect description of reality. But “use” is a rare experience with economic models, except forecasting models, and what we do instead is hang up the maps on our walls and discuss which is the most “beautiful”. Beauty might be a correlate of usefulness but it might not be.

CGE Modelling Helps to Make the Issues Clear

A general equilibrium effect that I think may be important for improvement in air quality in the LA basin is the induced change in residential location, away from the Coast toward the Interior as the relative air quality in the Interior is increased. This movement of residential location toward the Interior where air quality is relatively poor reduces the benefits from air quality control, perhaps a lot, perhaps not so

much. When we use the acronym DSGE, does that encompass residential location choices? I think not. This issue of locational endogeneity is explicitly discussed on page 46 of “Economy-Wide Modeling: Benefits of Air Quality Improvements White Paper” (September 22, 2015) which refers to CGE models which include it.

I am inclined to think that a huge benefit of the CGE approach is that it forces modelers to think about, to make and to reveal choices regarding what is important and what is not. I think the two documents “Economy-Wide Modeling: Benefits” and “Economy-Wide Modeling: Costs” are symptoms of a very healthy and broad conversation on this set of issues. The document “Economy-Wide Modeling: Benefits of Air Quality Improvements White Paper” has a sections on pages 17 and 24 titled “Potential Analytical Limitations” which are features-driven discussions of what’s excluded from certain CGE models. But reliability is important too. The reliability discussion is relegated mostly to a single paragraph on page 13:

“In addition to the limited uncertainty analysis focused on propagating standard errors from the epidemiological studies, benefits analyses typically include a number of sensitivity analyses for important analytical choices, such as alternative effect estimates, functional forms (e.g. threshold models), discount rates, lag structures, and inclusion of additional endpoints. While these sensitivity analyses are not formal uncertainty assessments, they provide insights into how sensitive overall benefit estimates are to different assumptions. In general, because of the large impact of PM_{2.5} on premature mortality and the large magnitude of the VSL, total benefits are most sensitive to assumptions regarding the PM_{2.5} health impact and valuation functions.”

I infer from this paragraph that reliability problems are fully understood by EPA personnel. They just don’t know how to handle them, while handling new features of a model is a relatively straightforward task. As for me, when I was working on the potential impacts of NAFTA, I read the first sections of CGE papers with great interest, since these had interesting and relevant ideas about the issues that needed to be understood, but I completely ignored the reams of numbers that came in the next sections. I had no reason to think they were reliable. Perhaps the way of saying it rhetorically is that a CGE model provides only a numerical theorem, a mapping of numerical assumptions into numerical implications. What would make us think that a numerical theorem accurately describes the impact of air quality regulations, even if some of its features are calibrated to actual data? What feature of the data supports the conclusion? How can we convince a doubter?

Unreliability comes from Statistical Uncertainty and Model Ambiguity

Conceptual frameworks and numerical evidence are combined to make models for policy purposes. Limited reliability comes from two sources: statistical uncertainty and model ambiguity. Statistical uncertainty, measured with standard errors and t-values, indicates limitations of the numerical evidence. Model ambiguity, measured with a sensitivity analysis, indicates the extent to which conclusions are conceptually fragile, meaning small changes in the assumptions lead to large changes in the conclusions.

Econometric Inference or CGE Calibration?

Both econometric estimation and CGE calibration rely on conceptual frameworks to turn data into information, and then into public policy advice. (Confession: I am a purveyor of the art of econometrics and look askance at the CGE approach. Here it comes.) The econometric approach has built-in automatic humility, since it is impossible to claim there is econometric evidence when the policy variable of interest doesn't have enough independent variability after controlling for the other important effects. Then the standard errors applicable to the public policy levers are reported to be large and the effects are said to be "statistically insignificant", not necessarily small but hard to measure. The art of econometrics is to choose a model with enough confounders/complexities that it is believable, but not so many confounders/complexities that they destroy the statistical accuracy of the estimate of the policy effect. A second phase of the art of econometrics is to perform a sensitivity analysis which shows how much the policy effect changes as the list of confounders is varied or other features of the model are changed. The hoped-for outcome occurs when the estimated policy effects are both statistically accurate and also sturdy (insensitive to perturbations in the model). This is pretty rare in economics.

I don't see either of these sources of unreliability prominent in the CGE literature with which I am familiar, and no automatic humility. On the contrary, CGE seems to be used when the econometric modelling has issued a loud humility alert, meaning that the historical evidence regarding the hypothetical policy is either nonexistent or painfully weak. I recognize that there are plenty of important public policy questions, especially ones that deal with systemic effects, for which the econometric approach cannot be pursued because there are no historically relevant cases to be studied. But then policy makers need to be alerted to the limits of knowledge.

AAA-H ratings for mortgage backed securities and for policy advice

Maybe we need a rating system when communicating policy advice. The system known as an aircraft is not certified as airworthy until extensive actual flight testing in calm and also turbulent conditions. Likewise, rather than AAA ratings, mortgage-backed securities based on sub-prime loans issued in 2004 and 2005 should have been rated AAA-H, where H stands for hypothetical, according to the model, but not tested in the turbulent conditions created by recessions. These AAA-H ratings would have prevented pension funds from acquiring mortgage-backed-securities and would have made the 2008/09 systemic collapse a lot more mild. More recently, large increases in minimum wages are being imposed by many jurisdictions, including Los Angeles. My advice to the City Council has been: be careful. While many minimum wage studies have found little evidence of adverse employment effects, the City of Los Angeles is raising the minimum wage to \$15 levels that will directly affect something like 35% of the workforce, an experiment for which there is no historical precedent. The evidence that was presented in support of this increase in the minimum wage might be ranked AAA-H, meaning that there is pretty clear evidence about small increments to the minimum wage but extending these to such a high minimum wage requires some serious guesswork. In this setting, the best solution is to increase the minimum wage slowly and call a halt when the costs exceed the benefits.

To be provocative, I suggest that EPA advice to Congress should come with one of several ratings: AAA, BBB, AAA-H or BBB-H. AAA means highly reliable evidence and no material sensitivity to assumptions. BBB-H means a weak evidential base and material sensitivity to hypotheticals. It strikes me that there are a lot of public policy decisions that have to be made in a BBB-H settings, meaning uncomfortably high amounts of both statistical uncertainty and considerable model ambiguity.

Inframarginal Value: Consumer and Producer Surplus

Last point: As an admitted outsider to the conversation about air-pollution controls, I am inclined to think that estimating changes in consumer and producer surplus are the most difficult tasks, since the behavioral trails that might reveal these surpluses are very faint, if they exist at all. A similar concern applies to using GDP as an indicator of aggregate welfare, since GDP is the *market* value of produced goods and services. Per standard economic theory, the market value is the marginal value, and all the inframarginal value (consumer and producer surplus) is ignored. Thus all those tech gadgets that have made your life so much better hardly show up in GDP because their market prices are so low. We do not have “full value” of production estimates because it is so hard to measure consumer and producer surplus and we make due with market values instead.

Locational decisions leave behavioral trails about the marginal values of air quality, but wise public policy really needs to include the full value – the consumer and producer surpluses. The EPA document “Memo on Using Other (Non-CGE) Economy-Wide Models to Estimate Social Cost of Air Regulation” (wrongly) dismisses all approaches but CGE and DSGE because, it is said, these other approaches cannot produce estimates of consumer and producer surplus:

“Several types of economy-wide models are discussed: input-output models, large-scale macro-econometric forecasting models, a hybrid between the two (input-output (I-O) macro-econometric models), and dynamic stochastic general equilibrium models (DSGE). It is our conclusion that, aside from CGE models, only DSGE models produce an estimate of changes in consumer and producer surplus required for benefit-cost analysis of policies. For this reason, the main portion of the memo discusses DSGE models while the other three types of economy-wide models are described in the appendix.”

If this is saying that large-scale macro-econometric forecasting models cannot produce estimates of changes in consumer and producer surplus, I agree. They were not designed for that purpose. But that is not the same as saying that the econometric estimation approach is incapable of estimating consumer and producer surplus when there are appropriate behavioral trails. In a simple supply and demand market model the consumer and producer surpluses depend on the shapes of the supply and demand curves. In a simple general equilibrium setting, the consumer and producer surpluses depend on the nonlinearities of marginal rates of substitution (consumer) and marginal rates of transformation (producer). It is of course possible to infer these effects from some observed data, though it may be difficult. But absence of evidence about the shapes of supply and demand curves, or the nonlinearities in marginal rates of substitution and transformation, doesn’t justify just making assumptions, and pretending that the knowledge is perfect.

On this subject, I am doubtful that CV surveys can reliably determine inframarginal values (WTP) and I do not know of any evidence that might support an opposite conclusion. Think about it first for something apparently familiar like an iPhone purchase and later for something unfamiliar like air quality improvements. When you buy a new iPhone for \$700 you reveal that its value to you is at least \$700. But what is the inframarginal value? What is the maximum amount you would pay for it? When you answer that question, please wipe from your mind the existence of a similar Samsung phone or any other smart phone. We are not interested in the familiar purchase decision which depends on the price of close substitutes. We want to know the total consumer value of smart phones. To answer the question, we need you to imagine yourself back in 1990 being offered an iPhone of today. How much would you have paid for it then? \$10,000? This is getting into a highly complex hypothetical conversation. You have no experience which allows you to answer this question.

