Focus on Data or Model? - A Discussion on Empirical Strategies

Wenlan Luo

Ex-postly, all empirical methods are developed to deal with imperfect data. We use Probit/Tobit only if we observe grouped data. We use IV/control-function only if we have endogeneity problems. We use non-parametric only if we don't know the exact function form. We use structural when we know nothing.

Though given the data, choosing a particular model is usually the most we can do to fix imperfections, I still feel that these methods introduce additional assumptions which complicate the problem. We need normalization of latent cutoffs and distribution assumptions to do Probit/Tobit. We need correct specification of control-function to restore the orthogonal condition; IV may be better, but still the exclusion assumption can only be argued but never be tested. The way we do local average for non-parametric is essentially to put some weight on the assumed continuity of function. And For structurals, we impose even more. I agree that the ultimate research style of Economics, as a Science, should be like Physics so that all observations can well fit in complete models. But up to now, without sufficient development in the biological basics of human being behaviors, I feel that the most safe way is to impose unverifiable assumptions as least as possible.

Ex-antely, the "data or model" question is more philosophical. It basically asks whether you'd like to spend 3 years on estimating a complicated model with imperfect data, or running an experiment/employing your resources to find a natural experiment with good data and run simple regression. (Robin said they spent 3 years estimating the model, and I conjecture it may take Angrist similar resources to get the data.) Therefore, choosing which data-problem pair does depend on Economists' taste.

However, if we are given the problem to address, there's additional tradeoff posed by data accessibility. Good data can only be collected covering limited areas, limited aspects of the problem or with limited samples. Therefore, the conclusion made with good data may be quite robust, but not so interesting. If we are to target on universal conclusion we need more representative data, but usually it's too costly to run very broad experiment

and there's usually no natural experiment happened to be. And therefore, it may be more efficient to save resources on perfect data generation but focus more on techniques to alleviate issues.

In summary, my point is that: (1) Given the data, choosing a particular method is usually the most we can do. We can't question so much on this though sophisticated models do introduce additional assumptions that may complicate the problem. (2) If we are free to choose problem, whether "data-intense" or "technique-intense" depends on economists' taste. Anyway, both strategies can address their corresponding audience. (3) If we are given the problem to solve, the tradeoff is that good data produces robust but very local result while universal data calls for techniques. Now I go through the three presentations in the labor week to illustrate my idea.

In Angrist's paper, they use randomization of enrollment to Charter school for excess applications as instrument to identify the return to Charter school versus normal schools. The data is good, IV is strong, implementation is complete and conclusion is convincing. But the result is less interesting. Firstly, the result is local in the sense that it only assesses the return of one kind of school in one state of US. Of course, as policy evaluation, the result advocates Charter school in Massachusetts stronger than any plain arguments. But we can say no more economics out of the case. Secondly, the result is local in the sense that it can't help figure out how much more investment in Charter School is enough. Reduced regression only reports the linear approximation of the local effect. I feel the absence of space for policy experiments is one of the most unpleasant parts of reduced form versus structural. Other things that I'm concerned about is the conditional independence assumption when assessing the effects of school characteristics on return may not be valid. There must be some reason why only urban school has "No-excuse" disciplines and therefore "No-excuse" should not directly explain the gap between urban/rural Charter school returns.

In Robin's paper, they build a job matching model with friction and productivity shocks, estimate parameters using employment survey data and experiment various policies with model. The problem is very interesting in the sense that they are to build the world from scratch and if it happens to match the real world well, we are confident to use the model to depict the real world. Recall this is exactly how scientists did physics in early ages. The powerfulness of structural here is that we do address some problem that reduced form can't handle. Firstly, for example in his presentation, he argued that reduced form underestimates sorting because there's occupation left vacant endogenously by low-productivity firm, and reduced form can not identify vacancy from data, but structural accounts for the unobserved behavior. Secondly, as I mentioned before, this is survey data thus very representative and the conclusion, if any, should be very general.

Thirdly, structural does allow policy experiments.

What I'm concerned about is: (1) The model is too complicated that there's no way for theoretical proof of identification. This was also discussed during the seminar. Though robust to initial values and it's a "safe" iterated procedure to solve steady state, I still worry about to what extent we can trust the estimation from the model if we can't be sure it's uniquely identified. (2) Continued with the point above, there are a lot of minor but maybe nontrivial assumptions in the story of the model (such as the asymmetric assumption that workers can search on job but firms can't replace existing workers). Lacking in the backgrounds of either labor or structural estimation, I don't know we need these assumptions because they are believed to be how the real world runs or just for simplification purpose; and how much do these assumptions affect the result. (3) The authors drop large part of the data for estimation purpose. Is it legitimate? Basically, my concerns can be summarized that if the purpose of the paper is to address empirical problems, but we sacrifice so much "accuracy" just serving modelling purpose, how much would this be better than just looking at the possibly biased reduced-form estimation. (Of course in Robin's problem, there's no parallel reduced-form to identify the matching process, but some partial version of the model may apply.)