Taylor & Francis
Taylor & Francis Group

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Discussion on " Transparency in Structural Research" by I. Andrews, M. Gentkow and J. Shapiro

Elie Tamer

To cite this article: Elie Tamer (2020) Discussion on " Transparency in Structural Research" by I. Andrews, M. Gentkow and J. Shapiro, Journal of Business & Economic Statistics, 38:4, 728-730, DOI: 10.1080/07350015.2020.1804917

To link to this article: https://doi.org/10.1080/07350015.2020.1804917

Published online: 28 Sep 2020.

Discussion on “Transparency in Structural Research” by I. Andrews, M. Gentkow, and J. Shapiro

Elie Tamer
Department of Economics, Harvard University, Cambridge, MA (*elietamer@fas.harvard.edu*)

1. SUMMARY

The question of sensitivity or robustness of empirical estimates from economic models is key to understanding the role various assumption play in forming empirical estimates from applied economic models. This is important because these estimates play a role in the conclusions that can be drawn for problems of evaluation of policies and also in problems of predicting the response to new policies. The accumulation of results in science requires transparency, how these results were obtained, and more importantly equipping the users of these results with enough tools to examine and critically analyze the assumptions used, the methods and the data. *Transparency* is a far reaching concept, but the article above tries and succeeds in framing its properties in terms of novel statistical metrics that are within reach. I applaud the researchers for taking us down this road. The article above, and other works by the authors' on similar topics have provided thoughtful new perspectives on the question of learning broadly from applied economic models in situations where some (researchers) may have a different view of what assumptions are harmless/harmful relative to a consumer of this research (reader). The article provides useful perspectives (and good ideas and practical procedures) on the question of whether and how communicating results based on empirical models is *transparent*. The key insight is that researcher and reader may have a differing priors on model pieces and so for the results to be transparent particular statistics should be communicated in empirical work.

Evidence from a study is convincing only if it is robust or less sensitive to adhoc (and often untestable) assumptions. One element in a discussion on robustness is that the various parties may not agree on what assumptions they consider worrisome. The starting point of my comment is models in which there is an agreement on the worrisome assumptions and so the question then becomes one of sensitivity or robustness to such assumptions. To this extent, I provide another complementary approach to the transparency analysis in the above article. This approach that I sketch (and it is a rather heuristic sketch with no attempt at rigor) provides a *mechanistic top down partial identification* view of transparency through the angle of global robustness and sensitivity. This approach is one where again the researcher can/does anticipate correctly what the reader (or the majority of them) is worried about, and then provides confidence regions (or CIs) that hold under a broader set of assumptions and thus are more likely to be acceptable. Usually, this approach sounds more like the classic semiparametric approach in econometrics. However, here, an issue that I believe handicapped serious applications of the semiparametrics literature is the mainstream view that treated identification in econometric models as a binary proposition: a parameter is either identified or not. And so to achieve semiparametric point identification one requires extra assumptions (typically large support and exclusion) and these sufficient point identification conditions are hard to verify in typical datasets. So, my proposal is to shed this binary view of identification once and for all, focus first on a list of assumptions that are possibly suspect and then provide theoretically valid CIs that hold when these assumptions are relaxed; these CIs would be valid whether or not the parameter of interest is point identified. Failures of point identification due to relaxing one or more assumptions is not a red flag but the size of the ensuing identified set may be.[1]

As a caveat, I will mention at the outset that the literature on robustness and sensitivity is wide ranging due to its importance in applied work and so naturally and deliberately I will miss various approaches and directions that one might think about such as robustness/sensitivity and decision making, Bayesian perspectives on robustness, and misspecified models, local robustness, etc. to name a few and will indeed only focus on how one can cast a particular solution to transparency in structural models as one of inference with partial identification.[2]

2. IDEA

We start with the economist's preferred model. The main question that the top down sensitivity approach is concerned with is the robustness of this model's conclusion—policy counterfactual, evaluation, prediction, decision, etc.—to various

[1]Failure of point identification due to a relaxation of some assumptions may not be worrisome if our set estimates are tiny.

[2]Much of the ideas in this discussion and the research program in general were part of the Cowles Lecture. See Tamer (2015).

Journal of Business & Economic Statistics
October 2020, Vol. 38, No. 4
DOI: 10.1080/07350015.2020.1804917

assumptions. This is complementary to the question asked in the AGS article but takes its starting point the parts of the model where potential *users* may find unconvincing. This robustness analysis that I seek is not sample dependent, but rather model dependent and it connects to various priors that researchers and users have. So, this approach takes the set of assumptions (functional form, distributional or behavioral assumptions) that researchers anticipate would be the most worrisome and weakens those without worrying about whether model parameters (or policy responses, etc.) are point identified. Two simple examples should help anchor our intuition.

Example 1 (Logit with discrete regressors). Consider the question of inference (estimation) on β in the binary response model[3]

$$y = 1[\alpha_0 + \alpha_1 x_1 + \beta x_2 \geq \epsilon] \tag{1}$$

with ϵ assumed to be statistically independent of $x = (x_1, x_2)$. In addition, assume that both x_1 and x_2 are binary 0/1 and data on (y, x) are available. This model does not make functional form restrictions on the distribution of ϵ but only requires that ϵ be independent of the covariates.[4] Now, suppose we estimate β via logit MLE for example and we get that $\beta = 0.5$. Is this estimate sensitive to the logit assumption? A researcher is worried that reporting the logit estimate might not be convincing to readers that are concerned about this distributional restriction.

We know from the semiparametrics literature (see Manski 1988; Komarova 2013) that with discrete support on the regressors, the model class above does not point identify $\theta = (\alpha_0, \alpha_1, \beta)$ (even though the logit likelihood has a unique maximum under the usual full rank conditions). So, then, what to make of the logit estimates? and how do we analyze this sensitivity? Note that a random sample identifies the cell probabilities $p_0 \equiv (P(y = 1|1,1), P(y = 1|(1,0), P(y = 1|(0,1), P(y = 1|(0,0))$. This p_0 can be mapped to the model implied choice probabilities $p(.;\theta, g)$ where we use our prior information on ϵ mainly that $\epsilon \perp x$ with a density $g \in \mathcal{G}$, a space of densities. What we are worried about as empirical researchers—and here the reason for this sensitivity/robustness analysis—is the possibility that there might be pairs (θ_1, g_1) and (θ_2, g_2) with $g_1, g_2 \in \mathcal{G}$ s.t.

$$p(.;\theta_1, g_1) = p(.;\theta_2, g_2) \equiv p_0.$$

And, the further apart θ_1 is from θ_2 the more sensitive (less robust) the model is to functional form assumptions on the distribution of ϵ. Indeed, the following set Θ_I summarizes this sensitivity

$$\Theta_I = \{\theta \in \Theta \subset \Re^3 : \quad p(.;\theta, g) = p_0 \quad \text{for some} \quad g \in \mathcal{G}\}.$$

This is the set that characterizes the sensitivity or robustness of the model to parametric assumptions on g. For instance, here with x_1 and x_2 binary, and if we observe all four realizations of (x_1, x_2): $(1,1), (0,1), (0,1)$, and $(0,0)$ it is easy to determine that in the base case[5] with $\alpha_0 = 0$, $\alpha_1 = 1$ and $\beta = 0.5$, we get that the sensitivity set for β is

$$\Theta_I(\beta) = [0, 1].$$

This is all we could learn about β under the model above (again, by model, I mean the threshold crossing model above with ϵ independent of x). For any b in $\Theta_I(\beta)$, we can find a density g for ϵ, such that $p_0 = p(.;(0, 1, b), g)$. This is a simple example where logit (or probit) estimates of β do not provide the full picture and the degree of the sensitivity depends on the magnitudes of the losses.

Finally, the logit likelihood is uniquely maximized at $\beta = 0.5$ and so $p(.;\beta = 0.5, \text{logit}) = p_0$ because for this simulation, the logit was well specified. It would be interesting to examine the relationship of the argmax of the logit likelihood to Θ_I.

This example casts global sensitivity/robustness as a top down partial identification problem. We start with a fully parametric model (logit discrete choice), and then ask how far would our estimates be when we relax a suspect assumptions *without worrying about point identification*. This is so because our inference methods ought to cover the identified set. This example illustrates this point as large covariate support is required to get point identification and we see that absent this, the model remains informative about β. If one is interested in partial effects, then it is possible to also define the identified set for partial effects in a similar manner. This is illustrated in Chen, Tamer, and Torgovitsky (2011) where methods for inference on partially identified semiparametric models are also provided and these methods can be used in this and other similar problems.

Example 2. A structural model of trade.[6]

Another approach to the top down sensitivity analysis is to look at the structural model of trade studied in Helpman, Melitz, and Rubinstein (2008) (HMR). The article examines the extensive margin of trade using a structural model estimated with current trade data. The following is a brief description of their empirical framework. Let M_{ij} denote the value of country i's imports from country j. This is only observed if country j exports to country i. If a random draw for productivity from country j to i is sufficiently high then j will export to i. To model this, HMR introduce a latent variable z_{ij}^* which measures trade volume between i and j. Here z_{ij}^* takes the value zero if j does not export to i and strictly positive otherwise. We adapt slightly their empirical model to obtain a selection model of the form:

$$\log M_{ij} = \begin{cases} \beta_0 + \lambda_j + \chi_i - \nu' d_{ij} + \delta z_{ij}^* + u_{ij} & \text{if } z_{ij^*} > 0 \\ \text{not observed} & \text{if } z_{ij^*} \leq 0 \end{cases}$$

$$z_{ij}^* = \beta_0^* + \lambda_j^* + \chi_i^* - \nu^{*\prime} d_{ij} + \eta_{ij}^*$$

in which λ_j, χ_i, λ_j^*, and χ_i^* are exporting and importing continent fixed effects, d_{ij} is a vector of observable trade frictions between i and j, and u_{ij} and η_{ij}^* are error terms described below.

[3]One can be interested in partial effect but we use β here for illustration.

[4]This assumption can be weakened.

[5]The semiparametric model requires a scale and location normalizations which would be achieved with $\alpha_0 = 0$ and $\alpha_1 = 1$.

[6]The econometric analysis of this trade model appeared in Chen, Christensen, and Tamer (2018) where general simulation based methods are provided for constructing CIs in models that are possibly partially identified.

Table 1. Model estimates and robust to heteroscedasticity CI: first row reports MLE estimates under homoscedasticity, second row reports CI under homoscedasticity, and third row reports CI that are robust to heteroscedasticity and partial identification.

Pref. model	Outcome equation
Distance	2.352
	[1.154,3.549]
	[0.242,0.509]
Border	−5.191
	[−7.077,−3.3404]
	[−2.611,−1.898]
Legal system	0.358
	[0.002,0.715]
	[−0.242,−0.009]
⋮	

NOTE: The last CIs were constructed using Procedure 3 in CTT.

Notice that the model is different from the usual Heckman selection model due to the presence of z_{ij}^* in the outcome equation. Exclusion restrictions can be imposed by setting one or several of the elements of ν equal to zero.

One issue (and here we could go in many directions to examine the sensitivity of this model) is that we may be worried about the impact of heteroscedasticity, a common feature of trade data. And unlike linear models, ignoring heteroscedasticity may lead to bias in nonlinear models. Also, with heteroscedasticity present, it is not clear that the model point identifies the parameters (without further restrictions on variation in terms of support conditions on the regressors). This is a similar problem to the one in Example 1 where we are worried that there may be θ_1 and θ_2 such that

$$p_0 \equiv p(.;\theta_1, g_1) = p(.;\theta_2, g_2),$$

where g_1 and g_2 are different heteroscedastic functions. To allow for heteroscedasticity, one way to proceed is to consider the model where the distribution of (u_{ij}, η_{ij}^*) conditional on observables is Gaussian with mean zero and covariance:

$$\Sigma(X_{ij}) = \begin{pmatrix} \sigma_m^2 & \rho\sigma_m\sigma_z(X_{ij}) \\ \rho\sigma_m\sigma_z(X_{ij}) & \sigma_z^2(X_{ij}) \end{pmatrix},$$

where

$$\sigma_z(X_{ij}) = \exp(\varpi_0 \log(\text{distance}) + \varpi_1 \log(\text{distance})^2).$$

More flexible parameterizations of the scedastic function are possible. Chen, Christensen, and Tamer (2018) provided approaches to likelihood based inference for parameters in the above model and obtained the estimates in the following Table.

Starting with the model above, we can estimate its parameter vector θ under homoscedasticity using MLE. Confidence intervals based on a t-statistic are provided in Table 1 underneath these estimates. For instance, the legal system variable has a parameter estimate of 0.358 with a CI of $[0.002, 0.715]$. This would be standard reporting in empirical articles. Here, I would advocate for including a "third row" below the usual CI. This is provided in the highlighted row and is a subvector CI on the scalar parameter that is robust to heteroscedasticity (and point identification). It is robust to the part of the model that the researcher anticipates a reader having the most problems with. As we can see, the marginal effect of this variable is negative according to a robust version of the model. Again, it is recommended that empirical articles report the third row in their tables (as opposed to the usual two) by adding a sensitivity/robust CI to the usual point and standard CIs to provide the reader with a view of how sensitive the estimates are to assumptions that are problematic. Of course, for this exercise, a researcher would need to anticipate the part of the model that causes the most unease.

3. CONCLUSION

It is my view that economic models are useful in collecting various evidences that policy makers can use to sharpen their priors. Economists are helpful in providing such information. However, economic models, as all models of science, are built on sets of assumptions, and some of these assumptions are not widely accepted. Given my belief that identification in economic models is NOT a binary question (a parameter is point identified or not), it is imperative for economists to report robust CIs that not only summarize statistical uncertainty, but also model uncertainty regardless of whether model parameters are identified. Model uncertainty is more critical than statistical uncertainty especially in structural models that are most useful for policy work. More broadly, the questions raised in this article above are immensely important and the approach it offers is also interesting and valuable. It adds to our arsenal of approaches to examining sensitivity of one's estimates and offers clarity in what and how various inputs lead to model outputs. The article and the ideas in it open up new approaches that would help make empirical economics research transparent, robust and hence more credible.

[Received June 2020. Revised June 2020.]

REFERENCES

Chen, X., Christensen, T. M., and Tamer, E. (2018), "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 86, 1965–2018. [729,730]

Chen, X., Tamer, E., and Torgovitsky, A. (2011), "Sensitivity Analysis in Semiparametric Likelihood Models," Cowles Foundation Discussion Paper No. 1836. [729]

Helpman, E., Melitz, M., and Rubinstein, Y. (2008), "Estimating Trade Flows: Trading Partners and Trading Volumes," *The Quarterly Journal of Economics*, 123, 441–487. [729]

Komarova, T. (2013), "Binary Choice Models With Discrete Regressors: Identification and Misspecification," *Journal of Econometrics*, 177, 14–33. [729]

Manski, C. F. (1988), "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83, 729–738. [729]

Tamer, E. (2015), "Sensitivity Analysis in Some Econometric Models," Cowles Lecture: ES World Congress, Montreal. [728]