Notes for Revision of "Mis-classified, Binary, Endogenous Regressors..."

Francis J. DiTraglia[1] and Camilo García-Jimeno[1,2]

[1]Department of Economics, University of Pennsylvania
[2]NBER

July 6, 2018

1 Comments from main editor

- "come up with an extremely short manuscript (about 10 pages for example) that focuses exclusively on identification."
  - This, I suppose, will mean that the paper should not be called "Mis-classified, Binary, Endogenous Regressors: Identification and Inference" anymore. Maybe something like "Identification in Models with Mis-classified, Binary, Endogenous Regressors"?
- "Your paper also needs improvements in the exposition".
  - OK.

Nothing else substantial from the main editor.

2 Comments from associate editor

- "Both referees disliked the inference part. I do not like it either, for reasons similar to the referees -especially what happens as $\alpha_0+\alpha_1 \to 1$. In light of this, I recommend that you remove completely the statistical inference part of the paper, and make this into a very short paper about identification, with a brief (≤ 1 page) discussion of inference and providing references to the existing procedures that can be applied.

  - I suppose we need to cut out all the inference section. However, I do think that in the letter we mention why we think that the limit as $\alpha_0+\alpha_1 \to 1$ is not interesting, and that the misclassification bounds will in general put you far away from that limit. We should also mention that besides this, the referees don't mention in their reports any other of the "reasons" why they dislike the inference.

- "The exposition in the paper is not fully satisfactory. You need to be more clear on what is completely new in this paper, and what was already known (including exact references to where the results had been derived)."

  - I suppose this means we mention who came up with the reverse regression bound, and the Bound/BBS/Mahajan stuff explicitly (for example in the first paragraph of page 8 in the current draft). I do not see what this has to do with "exposition".

- " You need to provide empirical examples for the various assumptions you propose -at least for the ones that one could call "classic" (e.g., your assumptions 2.1 and 2.2)."

  - AE suggests looking at a chapter in the HoE by Bound, Brown and Mathiowetz (2001) for such empirical examples. What this entails is (Ass. 1) finding examples of papers with additively separable errors, and valid binary instruments. I suppose the empirical examples do not need to acknowledge of deal with mis-classification, as we could just point out that the problem could exist in such empirical settings. Examples related to Ass. 2 are more about mis-classification, so we need some empirical examples where there is a binary misclassified regressor. I suppose the Kane, Rouse, Staiger (1999) example would be ok since Ass. 2 says nothing about endogeneity of the regressor. I saved the Bound, Brown and Mathiowetz (2001) chapter in the OLS vs. IV/References folder.

- "For the stronger assumptions that you impose to obtain point identification, I would recommend citing empirical papers that de-facto impose sufficient conditions."

  - Maybe look at papers that use heteroskedasticity type conditions? Lewbel etc...

- "I would also like you to address comment #1 from the referee whose report starts with "Reacting to an issue..." ". Comment #1 of Referee 1 is the following: "I believe that this paper will become useful if the authors can extend the current model to the one above and provide sharp partial identification results for typical treatment parameters such as the average treatment effects $E[\beta(x,\epsilon)|x]$ and $E[\beta(x,\epsilon)]$ for example. Otherwise

I would like to be convinced by the authors citing a list of empirical papers in eg AER that many empirical studies indeed consider the simplistic model $y = c(x) + \beta(x)T^* + \epsilon$ of binary treatments."

  - I suppose one thing to mention here is that even in the "simplistic" model there was NO CORRECT partial or point identification result in the literature.
  - I don't think that by "address" the comment, the AE expects us to derive results for a model with unobserved heterogeneity, but rather the alternative mentioned by the referee of citing papers that use the "simplistic" model. (Alejandro task?) I don't think these need to deal with mis-classification, only illustrate that some papers use binary endogenous regressors and don't necessarily talk about ATTs. The "many empirical studies" part of this seems more challenging.

- "I disagree with comment #3 of that referee, while I agree with his/her comment #2."

  - OK we will dis-regard comment 3 of Referee 1.
  - I do not understand comment 2 of Referee 1, particularly since right after Ass. 2 we say what 2.2 (i) means.

- "Theorem 2.1 is poorly stated, and similarly for its proof: -To begin, the theorem does not explicitly say for what parameter vector is the sharp identified set being characterized. I believe the vector is $[\alpha_0(x); \alpha_1(x); c(x); \beta(x)]$: Please clarify.

  - This is correct, we can explicitly say this.

- "-Next, Assumptions 2.1-(ii) and 2.2-(i) jointly yield a requirement on the observable conditional distribution of $T|z$; namely that $p_1(x) \neq p_0(x)$: If this condition fails in the data, the model is rejected and the sharp identification region should be empty. So Theorem 2.1 should include $p_1(x) \neq p_0(x)$ as an assumption.

  - I think the AE did not read the statement of the theorem correctly. The theorem BEGINS WITH THE STATEMENT: "Under Assumptions 2.1 and 2.2 (i)(ii)". These assumptions together IMPLY $p_1(x) \neq p_0(x)$, so I think it would be redundant to "include $p_1(x) \neq p_0(x)$ as an assumption." Unless I'm missing something.

- "-Concerning the proof, you should improve the explanation of what it is supposed to show... In the proof you want to show that if you build any probability distribution that satisfies the (in)equalities above, then all your assumptions are satisfied to. Am I

understanding your strategy correctly? If so, you should state this goal at the beginning of the proof."

  - YES, you are understanding correctly. We can certainly state this goal at the beginning of the proof.

- "Theorem 2.2 requires $E[y|x;T=0;z=k] \neq E[y|x;T=1;z=k]$: You state before the theorem that this holds "generically". Both referees complain about this assumption. I agree with them and would like you to show how it is related to your other assumptions. On the other hand, this assumption is testable so it might be worth mentioning it." Referee 1 refers to this in his comment (7), and Referee 2 refers to this in his comment (1).

  - My understanding is that the reason we argue that this holds generically, is that as long as the conditional means of y given T^*. are not equal, then the misclassification rates would have to satisfy a non-generic relationship for the conditional means given T to be the same. So I suppose we can derive exactly what this relationship would need to be.

- "Bound, Brown and Mathiowetz (2001) also discuss nicely examples where the non-differential measurement error assumption is warranted, and where it is not. Again a reference to them and some examples would benefit the paper."

  - OK. Mention of this probably should go near Theorem 2.2.

- "p. 12 contains two typos in the third line of the second paragraph, in the math expressions for the conditional expectations."

  - OK, although I only see one typo.

3 Comments from Referee 1

- "(1) The class of models considered in this paper seems to have limitations for practical use. Researchers using binary treatment models are almost always concerned about nonadditive unobservables entailing heterogeneous treatment effects that require models like $y = c(x,\epsilon) + \beta(x,\epsilon)T^*$ rather than $y = c(x) + \beta(x)T^* + \epsilon$ considered in the paper. Partial identification is a useful approach when researchers want to allow for more general models. These points taken into consideration I believe that this paper

will become useful if the authors can extend the current model to the one above and provide sharp partial identification results for typical treatment parameters such as the average treatment effects $E[\beta(x,\epsilon)|x]$ and $E[\beta(x,\epsilon)]$ for example. Otherwise I would like to be convinced by the authors citing a list of empirical papers in eg AER that many empirical studies indeed consider the simplistic model $y = c(x) + \beta(x)T^* + \epsilon$ of binary treatments."

  - Every class of models has some limitations.
  - See response to AE comment above.

- "(2) $\alpha_0(x,z)$ and $\alpha_1(x,z)$ are defined in (2) but neither $\alpha_0(x)$ nor $\alpha_1(x)$ seems to be defined before they are first used in Assumption 2.2 (i).

  - See response to AE comment above.

- "(3) The paper says Assumption 2.2 (ii) ... is equivalent to requiring that T and T? be positively correlated. The equivalent statement that "T and T^* be positively correlated is much clearer than the current statement "$\alpha_0(x) + \alpha_1(x) < 1$ of Assumption 2.2 (ii). Therefore I suggest that Assumption 2.2 (ii) be rewritten into the equivalent clearer statement."

  - AE asked us to ignore this comment.

- "the proof of Theorem 2.1 contains issues. (It may be my problems but a clearer proof will help.) Here is a nonexhaustive list of major ones. (i) The objective and the logic in the proof are unclear given what is stated in the theorem. I suggest that you clarify at the beginning of the proof what logical statements need to be shown to claim the sharp characterization and then proceed with proving them."

  - See response to AE above.

- "(ii) Related to item (i) above I do not necessarily doubt the claim of sharpness in Theorem 2.1 but I am afraid that the sharpness is not formally shown in the proof or I may be missing it."

  - Referee is missing it. Clarify.

- "(iii) The first two paragraphs of the proof of Theorem 2.1 attempt at showing and claiming what the authors call "valid probabilities. I assume that this term "valid probability refers to a valid probability measure. But none of the axioms of a probability measure other than the unit interval range is checked or invoked to show or claim that a function of interest is a "valid probability."

  - All we meant to say is that they should be numbers between 0 and 1.

- "(iv) $p_k^* \neq p_l^*$ is shown to hold provided that $p_k ? p_l \neq 0$ holds but I cannot find this latter condition stated in the theorem. I may be missing it but if this condition is indeed missing then what is claimed in the proof does not seem to hold."

  - I believe it is exactly the other way around. The assumption is for the stars, and those imply the non-stars.

- "(v) By Assumption 2.2 (ii) is mistaken for By Assumption 2.2 (i)."

  - Correct, this was a typo.

- "(vi) Case II says $p_k^* = 0$... requires $p_k = \alpha_0$. Please show calculations to claim this logical implication. Use other relevant assumptions if any is used here."

  - This is a direct implication of Lemma 2.1.

- "(vii) The arguments for Cases III and IV are to find conditional distribution functions that satisfy A.2, A.3 and Assumption 2.1 (iii). The proof is not clear about why these arguments prove what is claimed in the theorem."

  - OK, we can be more explicit about this.

- "(5) Corollary 2.2 is about what the authors consider as two common cases: one-sided misclassification and symmetric misclassification. I do not know how common they are indeed. It would help me if the authors can cite empirical validation studies where each of these two common cases is true."

  - Follow the suggestion from AE and look for examples in the handbook chapter.

- "6) It will help if the authors can more clearly explain their contributions to the literature for the partial identification results. Just before Theorem 2.1 it says Combining the two lemmas yields a well-known bound. Just before Theorem 2.2. it says the sharp

bounds that we derive by adding Assumption 2.2 (iii) are new to the literature. My understanding from these discussions is that Theorem 2.1 and their corollary rephrase what is known in the literature while Theorem 2.2 shows adding Assumption 2.2 (iii) to tighten the bounds is the contribution by the authors to the existing literature. However I am not certain about this understanding because the paper is not clear about this point. Please clearly write which ones of Theorem 2.1, Corollary 2.1, Corollary 2.2, Theorem 2.2, and Corollary 2.3 are first discovered by the present paper. For each of all the other theorems and corollaries please cite existing papers or textbooks that show the result."

  - The referee's understanding is incorrect: The result from Theorem 2.1 is new, as I don't think the identified set had been characterized before.
  - My understanding is that Corollary 2.1 is not new.
  - Theorem 2.2 is thus not the only contribution.
  - Corollary 2.2 is also new.
  - Theorem 2.2 is obviously new.
  - Corollary 2.3. is new since it is the actual statement of the lack of point identification.

- "(7) The additional condition $E[y|x, T = 0, z = k] \neq E[y|x, T = 1, z = k]$ used for Theorem 2.2 needs more explanation. Note that T is an outcome of measurement error and is not a primitive random variable. Provide a sufficient condition in terms of primitives for this high level condition."

  - See comment to AE above.

- "(8) A similar comment applies to the proof of Theorem 2.2 as the comment above for the proof of Theorem 2.1. Again I do not necessarily doubt the statement of the theorem as it is quite intuitive. It will help to have the proof written more clearly."

  - I suppose we can try to be more clear.

- "(9) The proof of Theorem 2.3 including the proofs of Lemmas 2.3 and 2.4 can be drastically simplified by skipping most of the arithmetic steps. I also point out a typo in the last part of the proof: ...strictly greater than zero since $\theta_1 \neq 0$ and $\alpha_0 + \alpha_1 \neq 0$ is mistaken for ...strictly greater than zero since $\theta_1 \neq 0$ and $\alpha_0 + \alpha_1 \neq 1$."

  - I suppose we can cut out arithmetic steps in the proof.
  - OK, typo should be corrected (p. 47).

- "(10) Suppressing the dependence on x is fine for the sake of simplicity of writing for identification results. On the other hand it is crucial for practical purpose that the implementation details for inference to include x. In other words I suggest that x be written in Section 3."

  - Since the inference section is going to go away, I suppose this can be ignored.

- "(11) The inference section discusses a weak identification problem. My understanding of weak identification in the GMM context is characterized by drifting population moment conditions or drifting population parameters (β in your case). Please clearly write such a characterization."

  - Since the inference section is going to go away, I suppose this can be ignored.

- "(12) I think the inference section can be to a large extent shrunk by resorting to the existing inference papers that you cite."

  - I disagree but since this is the suggestion of the AE, it's what we will do.

- "(13) The simulation study ignores covariates x. But the present model without the covariates is an extremely simple linear constant coefficient model $y = c + \beta T^* + \epsilon$ with binary regressor. To demonstrate the usefulness of the presented method it will be essential to run simulations with multiple covariates and to study how results vary across the dimension of x. The authors may want to refer to more recent projection methods in the set inference literature for presentations of set results under higher dimensions of x."

  - Clearly the referee did not read the footnote where we do refer to that recent literature.
  - Since the inference section is going to go, I suppose we can ignore this.

- "(14) I am curious about how empirical results for the proposed method and existing methods would look like for empirical data. It would be nice if the authors can demonstrate an application of the proposed method where the present model is suitable for the application. Demonstrate discrepancies or similarities among the OLS estimate, the IV estimate, estimates of the authors partial identification bounds, the estimate of the authors point identified estimand."

– Since the AE only wants a short paper with the identification results, we must ignore this suggestion.

4 Comments from Referee 2

- "... But the conclusion the authors come to is still not obvious to me even if I hold α_0 and α_1 fixed.

  – Unfortunately the referee does not mention why this is not obvious to him, so it is hard for us to address this.

- "... I am not convinced that similar difficulties do not arise when $\alpha_1 + \alpha_0$ is close to 1."

  – There are partial identification bounds for α_0 and α_1, so this is testable and in practice will always be ruled out by the data.

4.1 Major comments

- "1. How does the additional assumption that $E(y|x, T = 0, z = k) \neq E(y|x, T = 1, z = k)$ relate to all the other identifying assumptions? A brief explanation would be helpful here."

  – See response to associate editor on related point.

- "2... In estimation moment conditions arising both from point identification conditions and partial identification conditions are imposed. But I am not sure what exactly the point of this result is. In the end, estimation is done under assumptions 2.5 and 2.6. So are the bounds of Theorem 2.2 only informative when $\beta(x)$ is close to 0? "

  – No. Clearly the referee did not look at the simulation exercises, that SHOW that the non-differential assumption is informative for any value of β.

- "2 (cont.) ... Assumptions 2.5 and 2.6 are further restrictions on the conditional distribution of y. When $\beta(x)$ is close to 0, then would the bounds in Theorem 2.2 be tighter for $\alpha_1(x)$ and $\alpha_0(x)$ when assumptions 2.5 and 2.6 are imposed?"

  – I believe the answer is no, because 2nd and 3rd moments are informative about the α's only when θ_1 is not close to 0.

- "3. The proof of Theorem 2.3 starts with the statement that $\theta_1, \theta_2, \theta_3$ are identified, since $Cov(T, z) \neq 0$. Then the authors write ... so long as $\beta \neq 0$, we can rearrange Equations 7 and 8 to obtain.... I dont see why the qualifier at the beginning is needed. As far as I can tell, the only condition needed to write equations (A.8) and (A.9) using equations 7 and 8 is that $\theta_1 \neq 0$, which is already identified. The reason I am pointing this out is every other statement following this statement of the authors seem to have been made under this qualifier, which left me wondering what happens if that qualifier does not hold. But this qualifier seems to be unnecessary anyway."

  - It is true that all needed for eqs (A.8) and (A.9) is that $\theta_1 \neq 0$. But if $\theta_1 = 0$, which is implied by $\beta = 0$, then A.8) and (A.9) are not usable so the *alpha*'s cannot be learned. That is why we make the qualifier, I believe.

- "4. Towards the end of the first paragraph of Section 3, the authors write For simplicity we fix the exogenous covariates at some specified level and suppress the dependence o x in the notation. I do not see what assumption implies exogeneity of x. I also dont know why we need x to be exogenous. These thoughts led me to believe that when the authors talked about exogenous regressors they meant variables other than x. So calling x the vector of exogenous variables is misleading."

  - Since the inference stuff is going away, I suppose we can ignore this comment.

- "5. For inference, should I also be worried about the possibility that $\alpha_0 + \alpha_1$ can be arbitrarily close to 1? In $h_{Ik}(w_i, \theta, q)$, $1 - \alpha_1 - \alpha_0$ appears in the denominator. Does this fact create problems for the asymptotic variance of the estimators if $\alpha_1 + \alpha_0$ is close to 1?

  - The bounds on the α's make this not be a problem.

- "6. Related to the previous item: looking at the expression for β on page 14, the estimator does not exist when "

  - The comment appears incomplete so we do not know how to address it.

- "7. Lemma 3.1 is confusing. What does it mean to let $||h_n^I(\vartheta_0, g(\vartheta_0))|| \leq inf_q ||h_n^I(\vartheta, q)|| + o_P(1)$? Do the authors mean to let $\hat{q}$ be an estimator such that the statement holds? Also what is h_n^I?"

  - I believe the answer to the first question is yes.

- For the second question, in the statement if the lemma we say that h_n^I should be defined analogously to $\overline{m}_{1,n}^I(\vartheta)$.
- We may want this to be clarified for the "working paper version".

4.2 Minor comments

- "Top of p. 10, $(T = 1, z = k, T^* = 1)$ should be changed to $(T = t, z = k, T^* = 1)$."
  - OK.
- "Put "to between "not and "contain in Remark 2.1."
  - I cannot find any Remark 2.1 in the paper, not the words "not" and "contain" following each other...
- "Page 10: Replace ... all conditional densities have positive value bounded from 0 and ∞. by "... all conditional densities are bounded and bounded away from 0." "
  - I am unable to find this phrase either in page 10 or anywhere in the paper.