Report – Economics E-Journal – Replication Section

"Which Panel Data Estimator Should I Use?: A Corrigendum and Extension"

This paper studies the performance of different panel data estimators for situations with small to moderate numbers of cross-sectional units (N) and time periods (T) in the presence of cross-sectional dependence. It uses Monte-Carlo simulation to develop recommendations for applied researchers concerning the question which (built in in Stata or Eviews) estimator to use depending on the data and goal of the researcher (estimation or inference). The paper corrects and extends an earlier study by Reed and Ye (2011, Applied Economics, henceforth RY). The correction reflects a change in the parameter values used for the data generating processes (DGPs) that the authors claim better reflects real-world autocorrelation in explanatory variables. The extension consists of an inclusion of additional real-world datasets based on which the authors construct the DGPs' values. The correction leads to a simplification of RY's recommendations; the extension demonstrates that the recommendations also hold for different datasets and are thus more broadly applicable. However, the caveat remains that different estimators are recommended for estimation and for hypothesis testing. The authors further extend RY and suggest that using bootstrapping may help to allow researchers to use one estimator for both.

Overall, this is a nice, informative and clearly written paper that concerns an important topic: guidance on estimator selection. My main concern with the paper is that it does not fit the format suggested for the replication section as described in Economics E-Journal's Replication Guidelines. More specifically, the authors do not make an effort to actually reproduce and report RY's results. They only report the results based on the corrected DGPs. Even when putting the two papers side by side one cannot directly compare many of the results because they are not reported in the same way. [For example, Table 3 in RY reports efficiency separately for datasets with N<=T and for those with N>T, while Table 4 in the authors' new paper reports results for datasets with T/N>1.5 and T/N<=1.5] This makes it hard to judge how much the authors' correction to RY's DGPs affects the results. Given that this paper is submitted to be published in the replication section – and as I think one would actually learn something from directly comparing RY's results to the authors' results – the authors should include RY's results without the correction (in a way directly comparable to their results) in the paper.

In addition to some more minor comments that are listed below I have one additional major question: The authors report that their correction to how the DGP's values are constructed reduces autocorrelation in x, the explanatory variable, and thus reflects real-world datasets better. This correction leads to a simplification of RY's recommendations that now no longer have gaps (which depended on the level of heteroscedasticity) but apply more broadly. First, it would be nice if the authors provided some descriptive statistics of how much the autocorrelation in the constructed x changes due to their correction and how that better reflects real-world datasets (currently the autocorrelation in x is not reported in the paper). Second, the question arises whether RY's recommendations still hold for real-world datasets with high autocorrelation in x, i.e. whether in cases with high autocorrelation in x heteroscedasticity in the errors should be taken into account for estimator choice.

Minor comments and questions

1. Related to major comment 1: Why do the values in Table 2 for the RY datasets do not correspond exactly to the values in RY's Table 2? Shouldn't sigmas and rhos be the same if the same datasets were used to produce them?
2. The description of how the sigmas and rhos are constructed based on the real-world datasets is not entirely clear. Do the authors use the first N observations and then all contiguous T combinations for these observations or do they use all possible combinations of N cross-sectional units and with all contiguous T combinations?
3. While I like the authors idea to try to find a way to find a uniquely recommended estimator for both estimation and inference, I think the bootstrap exercise they present at the end of the paper does not really achieve this goal: The authors show that Estimator 7 (the recommended estimator for estimation for the reported cases of T/N combination) with bootstrapped standard errors outperforms Estimator 8 (the recommended estimator for inference). But can one be sure that Estimator 8 with bootstrapped standard errors would not perform even better than Estimator 7 with bootstrapped standard errors in terms of accuracy? They should at least note that based on this exercise they cannot uniquely recommend Estimator 7 but as this estimator's accuracy is very high with bootstrapped standard errors it is probably ok to use it.
4. The comment on IV in the conclusion seems a little out of place: I am not sure the authors should sell it as a contribution of their paper that they can make recommendations based on observed data characteristics while this is not possible in other situations. It would be a contribution if it had not been possible before to give recommendations based on observed characteristics for the problem that they pose themselves (i.e. cross-sectional dependence in panel data settings) – but it seems weird to sell it as a contribution when compared to entirely different problems.
5. Figures 1-8 (particularly 3 to 8) are hard to read in black and white print. It may help to introduce different shapes for the different lines that the reader is supposed to look at (and combine this with color coding if necessary).
6. The paper should be spell-checked and checked for duplicate words and punctuation errors.