Referee report on

‘Replication to assess statistical adequacy’.

The paper proposes a method for replicating empirical studies that focuses specifically on the statistical adequacy of the published models. It argues that often misspecification tests are not conducted or reported, questioning the basis on which published papers claim significant empirical results. By checking the statistical adequacy of the models, the replication exercise can confirm or question the reported results. The paper demonstrates how this would be done using the study by Acemoglu, Gallego, and Robinson (2014) as an example.

I fully agree with the premise of the paper. The author outlines the case for replication to specifically address the question of statistical adequacy clearly. The proposal is an example of pure replication, as outlined by Hamermesh (2007), directly assessing a published study’s quantitative and qualitative findings. In an ideal world this would never be required. It should be incumbent on the authors of the original study to compute and publish misspecification tests relating to their models. If this was lacking, it should be incumbent on the referees to require the authors to include the results of the relevant misspecification tests prior to a journal accepting the paper for publication. This should be best practice, similar to journals now often requiring data and code to be submitted with the paper. To be at a stage where misspecification tests are not routinely included in applied papers is a poor reflection on our discipline, and the replication proposal would go some way to exposing this flaw.

I don’t have any substantive criticisms of the paper, but I make some more general comments below.

Comments

1. There is variation in the consequences of misspecification for the properties of estimators and tests. As specification in itself is not an end goal of an empirical study, but is required to validate the quantitative results claimed, then the nuances of misspecification failure are actually relevant. For example, some forms of non-normality may lead to unknown t-statistic distributions in small samples, but could be addressed by robust inference techniques or estimation methods that are robust to outliers. But autocorrelated disturbances in a dynamic model would lead to inconsistent parameter estimates, which is much more serious.

   One suggestion would be to extend the replication study, first undertaking the relevant misspecification tests as a ‘pure replication’ and then proceeding to find a congruent or well-specified model using the same dataset as ‘scientific replication’ in Hamermesh’s terminology. If the parameter estimates did not vary significantly then, although the initial study is invalid, its interpretation may not be. Of course, finding a vastly different congruent model would invalidate the initial study results further.

2. What the misspecification tests aim to address is whether the model is congruent or not, i.e. does it capture the characteristics of the unknown Data Generating Process (DGP), see Bontemps and Mizon (2003). The researcher needs to define congruency in the relevant context. This is clearly done in the paper for the AGR study discussed. But it is worth emphasizing that there isn't a 'one size fits all' set of misspecification tests that apply to all empirical papers. There is a trade-off, as with any statistical testing. Sufficient tests are needed to ensure congruency but they come at a price, as more tests increase the probability of rejection under the null. The tests must have the correct size properties and sufficient power when the relevant null hypothesis is false. Section 5 briefly mentions multiple testing but this is at the heart of the choice of misspecification tests.

3. The necessary tests for statistical adequacy will vary depending on the purpose of the model. If the aim is to test theory or say something about the parameters of interest, then weak exogeneity is required, and a statistically adequate specification as outlined on page 8 suffices. If the purpose is conditional forecasting, then the model also requires Granger non-causality, or strong exogeneity. And if the purpose is for policy analysis, then parameter invariance of the conditional model to interventions in the marginal model is also needed, i.e. super exogeneity. These are testable assumptions, so would fit into the misspecification framework, see Hendry (1995, ch.5).

4. A useful reference is Stigum (2014) in which he provides an approach to confronting theory with the data. This is similar to the Spanos approach outlined in the paper and provides motivation for the underlying argument of statistical adequacy.

5. The paper makes clear the general principles of misspecification testing and those tests that are relevant to the AGR study discussed, focusing on relevant tests for OLS and IV (2SLS). Some comments on the relevant tests for GMM and MLE given their statistical assumptions would be helpful for readers in the general discussion.

References

Acemoglu, D., F. A. Gallego, and J. A. Robinson (2014). Institutions, human capital, and development. *Annual Review of Economics 6*, 875–912.

Bontemps, C. and G. E. Mizon (2003). Congruence and encompassing. In B. P. Stigum (Ed.), *Econometrics and the Philosophy of Economics*, pp. 354–378. Princeton: Princeton University Press.

Hamermesh, D. S. (2007). Replication in economics. Working Paper 13026, National Bureau of Economic Research.

Hendry, D. F. (1995). *Dynamic Econometrics.* Oxford: Oxford University Press.

Stigum, B. P. (2014). *Econometrics in a Formal Science of Economics. Theory and the Measurement of Economic Relations.* Cambridge, Massachusetts: MIT Press.