Discussion of "Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models" by Søren Johansen and Bent Nielsen

Jurgen A. Doornik and David F. Hendry*
Economics Department and Institute for New Economic Thinking
at the Oxford Martin School, University of Oxford, UK.

October 2, 2015

1 Model selection allowing for outliers and shifts

We are delighted to contribute to the discussion of this paper by Søren Johansen and Bent Nielsen (JN henceforth) and their important development of an asymptotic theory applicable to a range of outlier detection algorithms.

The robust literature has two ways of handling outliers: using a soft or a hard weighting scheme. In the former, observations are given a weight between zero and one, while in the latter outlying observations are identified and removed from the estimation sample. In empirical modelling there is often a preference for identifying outliers, because they may be associated with a particular event.

In linear regression models, outliers can be removed by adding the corresponding impulse indicator variables to the model. This implies that identifying outliers amounts to selecting impulse dummies, providing a bridge between model selection and robust estimation. The model can be saturated by all possible impulse dummies, followed by selection of the ones that matter. This is impulse indicator saturation (IIS).

Johansen and Nielsen have a series of joint papers on this topic, starting from the development of an asymptotic theory of IIS in Johansen and Nielsen (2009). The current paper is another important contribution because the authors:

1. develop powerful tools to analyze the asymptotic behaviour of *iterative* procedures,
2. analyze more general settings that are relevant for *dynamic* models,
3. develop asymptotic theory for the *gauge*.

Treating outliers as a model selection problem has two advantages:

1. identification of outliers and selection of variables can be done jointly,

*Financial support from the Open Society Foundations and the Oxford Martin School is gratefully acknowledged.
email: jurgen.doornik@nuffield.ox.ac.uk and david.hendry@nuffield.ox.ac.uk

2. different patterns can be considered, such as structural breaks in macro-economic models captured by step impulse saturation (SIS, Castle, Doornik, Hendry, and Pretis, 2015) or volcano 'signatures' in climate models as in Pretis, Schneider, Smerdon, and Hendry (2015).

This in turn requires a model selection procedure that can handle more variables than observations and has good operational properties in a wide range of settings. *Autometrics*, see Doornik (2009), is such a procedure, using a general-to-specific approach that is described in some detail in Hendry and Doornik (2014).

Most of the statistical analysis of *Autometrics* so far has been using Monte Carlo analysis, focussing on the gauge and potency to express the success of model selection, see Hendry and Doornik (2014). It is therefore exciting to see the progress Johansen and Nielsen are making with studying the asymptotics of iterative procedures and gauge.

2 Illustration: returning to the roots of IIS

The accidental discovery of IIS occurred when analyzing Tobin's food data in Hendry (1999), so we return to that setting to illustrate its application. Here we consider another aspect of the data, namely the US savings rate, s_t, and its relation to the real short-term interest rate, Rr_t. The time series of s_t (solid line) and Rr_t (dotted line) are shown in Figure 1, and its relation to Rr_t in Figure 2.[1] The impact on saving of the regime shift to war-time rationing is stark, and a failure to take account of the resulting 'forced savings' could distort empirical analyses.

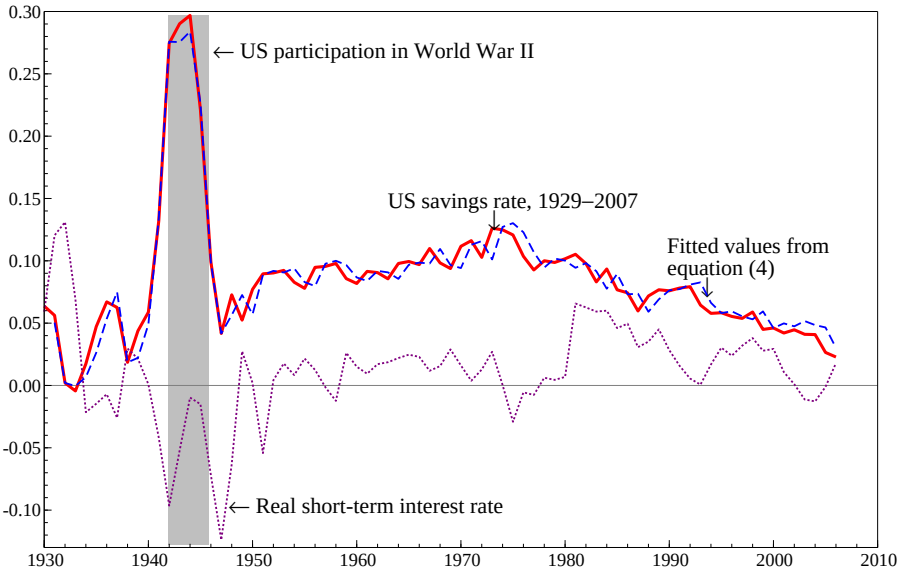

Figure 1: US savings rate and real interest rates, 1930 − 2006

[1]The calculations use OxMetrics, see Hendry and Doornik (2013), and Ox 7.1, see Doornik (2013).

The bivariate ordinary least squares (OLS) regression of s_t on Rr_t estimated over 1931–2006 is:

$$\widehat{s}_t = 0.091 - 0.50 Rr_t. \tag{1}$$

Applying m-step robustified OLS at $\alpha = 0.01$ first finds 1942, 1943 and 1944 as outliers. Adding impulse dummies for these observations and re-estimating finds 1945, after which there are no more:

$$\underset{(s.e.)}{\widehat{s}_t} = \underset{(0.004)}{0.077} - \underset{(0.095)}{0.143} Rr_t + \underset{(0.03)}{0.18}\, 1_{42} + \underset{(0.03)}{0.21}\, 1_{43} + \underset{(0.03)}{0.22}\, 1_{44} + \underset{(0.03)}{0.14}\, 1_{45}, \tag{2}$$

$$\widehat{\sigma} = 0.029 \;\; \mathsf{R}^2 = 0.71 \;\; t = 1931, ..., 2006.$$

The regression line from (2) is given by:

$$\widehat{s}_t = 0.077 - 0.14 Rr_t. \tag{3}$$

These two bivariate regression lines are shown in Figure 2, where the war-time regime outcomes captured by the outliers are shown in the ellipse and the first few years of the Great Depression in the bottom-right rectangle. Relative to the OLS line of (1), 1945 and 1947 are about equidistant. However, after removing 1942–44, the regression line tilts towards the horizontal, and 1945 becomes an outlier; 1932 and 1933 become more anomalous. Despite the huge rise in US unemployment after 1929, the real interest rate remained at very high levels, undoubtedly exacerbating the depression and leading to dis-saving.

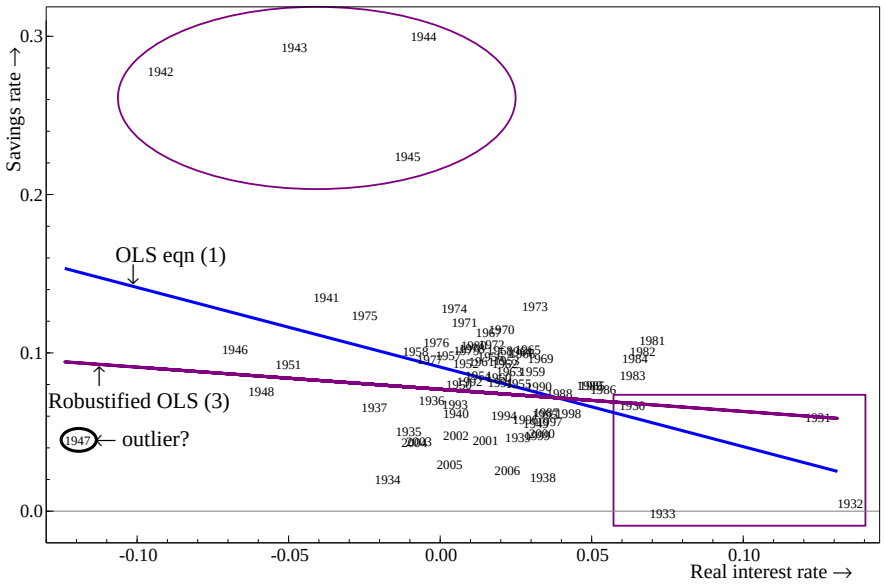

Figure 2: Impact of IIS removing outliers in cross plot of savings and interest rates

There are some serious issues. We did not apply the consistency factor of (JN 6.8), because at $c = 2.576$ it is close to one (JN Fig. 3). The results are probably not obtained under the null

anyway. We also ignored congruence. In (1) all mis-specification tests (c.f. JN eqn. 2.1) strongly reject, except for the RESET test. But when there are outliers, congruence is rejected anyway. However, (2) remains problematic: the test for serial correlation rejects strongly. Removing the four outlying observations is not enough to obtain a congruent representation.

Ignored outliers can distort inference, so need detection, but methods for doing so must be integral to model selection across potential candidate variables, possibly non-linearities, lag lengths, and both outliers and location shifts while retaining focus parameters. There are four fundamental problems facing empirical modelling:

1. a sufficiently general initial model must be formulated to capture all substantively relevant influences;

2. such a formulation will often have more variables than observations, so a powerful model selection procedure is essential, with tight significance levels matching its gauge;

3. any claimed relationships must be rigorously evaluated for congruence;

4. the selection algorithm must check immense numbers of possible models in a reasonable time.

A dynamic model of the relation between s_t and Rr_t would allow for lagged reactions. Without prior knowledge of the dynamic response, we allow for lag length up to three: past savings accumulate as wealth. Using *Autometrics* to combine IIS with model selection at $\alpha = 0.01$ and forcing the contemporaneous Rr_t, finds six outliers (1938,41,42,45–47). The only lagged variable surviving is s_{t-1}. The four war period dummies add up to zero (with a p-value of 57%), and we create a W_t variable to capture this period. From 1941 to 1946, W_t equals $0.5, 1, 0, 0, -0.5, -1$ respectively, and zero otherwise. The initial lag length of three reduces the estimation sample to 1933 – 2006; moving the start back to 1931 shows 1932 as an outlier by inspecting the residuals. The selected model is:

$$\widehat{s}_t = \underset{(0.003)}{0.008} - \underset{(0.04)}{0.15} Rr_t + \underset{(0.03)}{0.95} s_{t-1} - \underset{(0.01)}{0.04}\, 1_{32} - \underset{(0.01)}{0.04}\, 1_{38} + \underset{(0.007)}{0.13} W_t - \underset{(0.01)}{0.08}\, 1_{47}, \qquad (4)$$

$$\widehat{\sigma} = 0.011 \;\; \mathsf{R}^2 = 0.96 \;\; t = 1931, ..., 2006.$$

There are no significant mis-specification tests in (4), and $\widehat{\sigma} = 0.011$ which is less than half that in (2). With dynamics the pattern of outliers is changed, but the coefficient on Rr_t is almost identical to that in (2). There is no cointegration between s_t and Rr_t, probably due to the absence of other substantively important variables like unemployment or income.

Next, we apply the forward search (FS) to the general model from which (4) was found. We adopt $\alpha = 0.3$: $\psi_0 = \psi_1 = 0.7$, use a gauge of $\gamma = 0.01$, and LTS with breakdown ψ_0 as the initial estimator.

Figure 3 shows the FS path for the general model in the line marked with triangles. This is almost entirely outside the selected gauge, resulting in the detection of 22 outliers, which seems too many. If we only include lags up to one (line with circles), the minimum required to nest (4), we still have a similar unsatisfactory behaviour. Applying FS to the static model finds (2) as well as 1933.

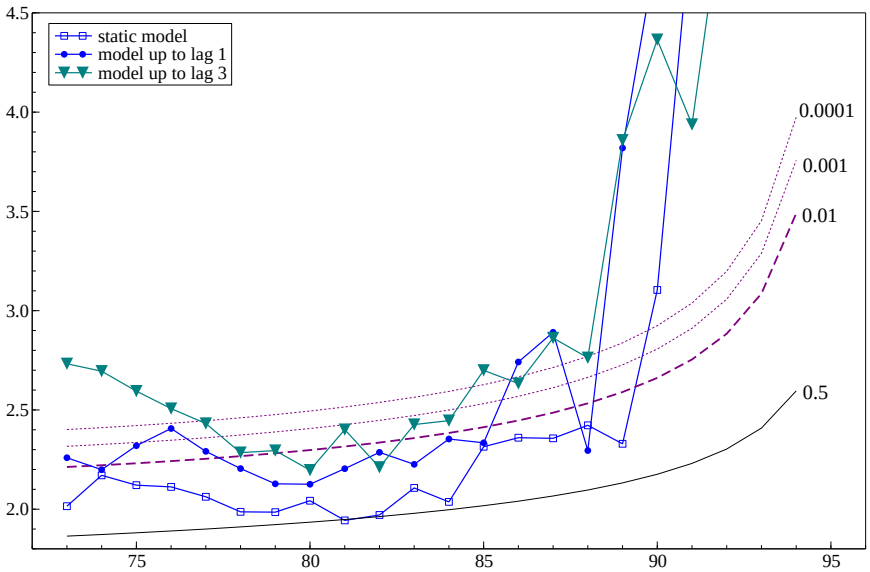

Figure 3: Forward squared residuals for US savings rate using $\psi_0 = \psi_1 = 0.7$. The smooth curves are the median and pointwise exit bands for several gauges.

3 Implications

Hendry and Mizon (2014) explain why it is essential to address shifts of distributions in empirical macro-econometric modelling, and not simply impose pre-conceived specifications. Much empirical analysis in macroeconomics concerns drawing policy implications, and these often entail location shifts (e.g., interest rates rising from a mean of 2% to say 5%) and occur facing other location shifts (e.g., price inflation jumping from 4% to 7%). Consequently, a key use of IIS and SIS is to test the invariance of relationships to shifts in their conditioning variables, a concept known as super exogenity: see Engle, Hendry, and Richard (1983), Hendry and Santos (2010). There are many large indicators in (4), but only 1_{1932} is in common with the final model of food demand in Hendry and Mizon (2011) (denoted HM). In their equation, 1_{1932} was due to a food program, whereas in (4) it is probably due to the huge rise in US unemployment eroding incomes. Importantly, the other indicators which substantively shift s_t do not occur in HM, whereas other indicators in HM do not enter (4): they are jointly insignificant in a test for omitted variables. If HM had simply been a regression, rather than a causal relation, the coefficient of s_t would not have remained constant after the large shifts in that variable shown by (4), so IIS helps establish the super exogeneity of s_t for food demand where it has an almost 1-1 long-run effect on food demand.

4 Conclusions

Johansen and Nielsen have provided important steps forward in the analysis of robust statistical methods, several of which are pertinent to our approach. A formal analysis of the statistical properties of iterative methods for jointly selecting variables, lags, non-linearities, outliers and shifts remains to be developed, but their general approach holds considerable promise.

References

Castle, J. L., J. A. Doornik, D. F. Hendry, and F. Pretis (2015). Detecting location shifts during model selection by step-indicator saturation. *Econometrics 3(2)*, 240–264.

Castle, J. L. and N. Shephard (Eds.) (2009). *The Methodology and Practice of Econometrics*. Oxford: Oxford University Press.

Doornik, J. A. (2009). Autometrics. See Castle and Shephard (2009), pp. 88–121.

Doornik, J. A. (2013). *Object-Oriented Matrix Programming using Ox* (7th ed.). London: Timberlake Consultants Press.

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983). Exogeneity. *Econometrica 51*, 277–304.

Hendry, D. F. (1999). An econometric analysis of US food expenditure, 1931–1989. In J. R. Magnus and M. S. Morgan (Eds.), *Methodology and Tacit Knowledge: Two Experiments in Econometrics*, pp. 341–361. Chichester: John Wiley and Sons.

Hendry, D. F. and J. A. Doornik (2013). *Empirical Econometric Modelling using PcGive: Volume I* (7th ed.). London: Timberlake Consultants Press.

Hendry, D. F. and J. A. Doornik (2014). *Empirical Model Discovery and Theory Evaluation*. Cambridge, Mass.: MIT Press.

Hendry, D. F. and G. E. Mizon (2011). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics 3 (1)*, DOI: 10.2202/1941–1928.1100.

Hendry, D. F. and G. E. Mizon (2014). Unpredictability in economic analysis, econometric modeling and forecasting. *Journal of Econometrics 182*, 186–195. Also see `www.voxeu.org/article/why-standard-macro-models-fail-crises`.

Hendry, D. F. and C. Santos (2010). An automatic test of super exogeneity. In M. W. Watson, T. Bollerslev, and J. Russell (Eds.), *Volatility and Time Series Econometrics*, pp. 164–193. Oxford: Oxford University Press.

Johansen, S. and B. Nielsen (2009). An analysis of the indicator saturation estimator as a robust regression estimator. See Castle and Shephard (2009), pp. 1–36.

Pretis, F., L. Schneider, J. E. Smerdon, and D. F. Hendry (2015). Detecting volcanic impacts in temperature reconstructions by designed break-indicator saturation. Working paper, Economics Department, Oxford University.