Noname manuscript No.
(will be inserted by the editor)

How to use SETAR models in gretl

Federico Lampis · Ignacio Diaz-Emparanza · Anindya Banerjee

Received: date / Accepted: date

Abstract This paper presents a means for the diffusion of the Self-Exciting Threshold Autoregressive (SETAR) model. Based on the Hansen (2000) methodology, we implement a function in **gretl** with which estimate a SETAR model. The function is provided with a nice graphical user interface that enables the average user to estimate a SETAR model and make inference easily. The function and its use is presented by means of a case study. In addition we show more functionalities of **gretl** in order to perform a preliminary analysis of the data.

Keywords SETAR models · free and open-source software · **gretl**

1 Introduction

The most famous procedure to estimate a Self-Exciting Threshold Autoregressive (SETAR) model is that of Tong (1990), while the most common approach to testing and making inference is due to Chan (1993). A part the main approach of Tong and Chan, the method of Hansen (1996, 1997, 2000) represents the most interesting alternative. Indeed Hansen proposes a confidence interval for the threshold parameter and a bootstrap method to test the linearity of the model. In addition, Hansen states that the standard confidence intervals for the autoregressive coefficients of the SETAR model could be not correct in small samples and proposes another method to calculate them. The purpose

Federico Lampis
University of Birmingham,
E-mail: f.lampis@bham.ac.uk

Ignacio Diaz-Emparanza
University of the Basque Country
E-mail: ignacio.diaz-emparanza@ehu.es

Anindya Banerjee
University of Birmingham,
E-mail: a.banerjee@bham.ac.uk

of this paper is to make the Hansen's procedures widely available and contribute to the diffusion of the SETAR models, both in research and teaching. Starting with the original Hansen code we develop a graphical user interface (GUI) in **gretl** with which estimate a SETAR model and make inference. Our work is another evidence of the use of **gretl** to disseminate and promote new econometric models; indeed joint to the explanation on how to use the GUI we present some additional functionalities of the program with which make a preliminary analysis previous to the estimation of the SETAR model.

gretl is a free and open-source software, see Cottrell and Lucchetti (2011) for an overview on the software. But among the various Free and Open-Source Software econometrics packages, **gretl** has the advantage of having an excellent user interface for the average user, similar to other popular commercial programs. As Smith and Mixon (2006) highlight, this GUI is the major reason that this software is very useful for teaching. There are already many econometrics books and courses based on **gretl**, among them see Adkins (2010). Its easy use makes it advantageous and profitable for academic institutions. Another important characteristic of **gretl** that makes it suited for research purposes is its numerical accuracy, which is as good as or better than other commercial programs, as shown by Yalta and Yalta (2007) and Baiocchi and Distaso (2003). All these characteristics make **gretl** a very powerful tool in research, and an increasing number of econometricians have begun to write their own code directly in **gretl**. For our purposes, the most outstanding feature of **gretl** is its clear and easy programming language (**hansl**) that allows us to write many statistical and econometrics operations with few commands.

The user can write code and run it in a command-line client (CLI also called **gretl** console). Moreover, this code can be easily packaged as a function, provided that it uses a GUI and is distributed to the scientific community which is exactly what we do with our code. Our function is freely downloadable from the **gretl** server (`http://gretl.sourceforge.net/`). Any user can study the code and search for bugs or, if necessary, modify the code. In this manner, our work can specifically contribute to the use of the SETAR models and the Hansen methodology. In the next section present briefly the SETAR model and the inference under Hansen's approach. In the fourth section we present a typical case of study and how to use **gretl** to make a preliminary analysis of the data. Finally we explain the management of the GUI function, and how to estimate a SETAR model and make inferences based on this procedure.

2 The SETAR model and inference under Hansen's approach

A SETAR with two regimes is defined as:

$$\begin{aligned} y_t = &(\alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p}) I(y_t \leq \gamma) + \\ &(\beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p}) I(y_t > \gamma) + u_t \end{aligned} \tag{1}$$

where y_{t-d} is the threshold variable that defines which regime is operating at time t and the error u_t is assumed to be $iid(0, \sigma^2)$. The change from one regime to the other is determined by the indicator function $I(\cdot)$, p is the

autoregressive order of the model, γ is the threshold parameter and d is known as the delay parameter. The parameters α_j are the autoregressive coefficients of the **lower regime** (observations in which $y_{t-d} \leq \gamma$), and β_j are the coefficients of the **upper regime** (observations in which $y_{t-d} > \gamma$). The equation (1) can be written in a more compact form as:

$$y_t = \theta x_t(\gamma)' + u_t \tag{2}$$

where $\theta = (\alpha'\ \beta')' = \big((\alpha_0\ \alpha_1\ \cdots\ \alpha_p)\ (\beta_0\ \beta_1\ \cdots\ \beta_p)\big)'$ and $x_t(\gamma)' = [x'_t I(y_{t-d} \leq \gamma)\ x'_t I\ (y_{t-d} \leq \gamma)]$. Once the SETAR model is reduced to equation (2) for a given value of γ, the Conditional Least Square estimator of $\theta(\gamma)$ is:

$$\hat{\theta}(\gamma) = \left[\sum_{t=1}^{n} x_t(\gamma)x_t(\gamma)'\right]^{-1}\left[\sum_{t=1}^{n} x_t(\gamma)y_t\right] \tag{3}$$

The residuals of this model are $\hat{u} = y_t - x_t(\gamma)'\hat{\theta}(\gamma)$ and for estimating their variance, we may use: $\hat{\sigma}^2(\gamma) = \frac{1}{n}\sum_{t=1}^{n} \hat{u}'_t(\gamma)$ When γ is unknown it has to be estimated minimizing the residual variance then the estimators for $\hat{\theta} = \hat{\theta}(\hat{\gamma})$ and $\hat{\sigma}^2 = \hat{\sigma}^2(\hat{\gamma})$ are obtained.

The great advantage of Hansen's methodology with respect to other approaches is the possibility to make inference on the threshold parameter γ. Hansen (1997, 2000) suggests using a likelihood ratio statistic to check for a specified value of the γ parameter under the null, the test being:

$$LR(\gamma_0) = n\left(\frac{\hat{\sigma}^2(\gamma_0) - \hat{\sigma}^2(\hat{\gamma})}{\hat{\sigma}^2(\hat{\gamma})}\right), \quad \text{with} \quad \begin{cases} H_0 : \gamma = \gamma_0 \\ H_1 : \gamma \neq \gamma_0 \end{cases} \tag{4}$$

The confidence interval for $\hat{\gamma}$ is done by the set $\widehat{\Gamma}$ of all values of γ for which the H_0 of the (4) is not rejected, at a significance level $(1-\xi)\%$: $\widehat{\Gamma} = \{\gamma : LR(\gamma) \leq c(\xi)\}$, where $c(\xi)$ is the ξ percentile of the asymptotic distribution of the statistic LR. The **linearity test** can be set out as

$$F(\hat{\gamma}) = n\left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}\right), \quad \text{with} \quad \begin{cases} H_0 : \beta = \alpha \rightarrow y_t = \beta x_t{}' + u_t \\ H_1 : \beta \neq \alpha \rightarrow y_t = \theta x_t(\gamma)' + u_t \end{cases} \tag{5}$$

where $\tilde{\sigma}^2$ is the variance of linear AR(p) under the null and $\hat{\sigma}^2$ is the variance of the SETAR model. The use of this test is hampared by the presence of nuisance parameter; since γ is not identified under the null hypothesis, the asymptotic distribution of $F(\hat{\gamma})$ is not a χ^2. Hansen (1996) proposes to approximate the asymptotic distribution of $F(\hat{\gamma})$ by means of bootstrap methods and then calculate the critical values of the test. Hansen (1997) also suggests an alternative methodology to calculate the confidence intervals for θ: **i)** for another significance level $\phi < 1$, the confidence interval $\widehat{\Gamma}_\phi = \{\gamma : LR(\gamma) \leq c(\phi)\}$ is computed, **ii)** for each $\gamma \in \hat{\Gamma}_\phi$ the confidence interval for θ is calculated according the standard asymptotic theory, **iii)** finally the new confidence interval $\hat{\Theta}_\phi$ is obtained by the union of these intervals. All the inference approach proposed by Hansen still holds in case of Heteroskedasticity[1] .

[1] The the error term u_t is assumed to be a heteroskedasitc Martingale difference sequence with respect to the past history of y_t.

3 A typical case of study

A very famous time series in the nonlinear analysis literature, the annual number of Lynx trapped in the Mackenzie River district in northwestern Canada, will serve to explain the use and results of the **gretl** SETAR function package. Tong and Lim (1980) and Tong (1983, 1990) proposed the use of a SETAR model for this series[2]:

$$y_t = \begin{cases} 0.62 + 1.25y_{t-1} - 0.43y_{t-2} + u_t^{(1)} & y_{t-2} \leq 3.25 \\ 2.25 + 1.52y_{t-1} - 1.24y_{t-2} + u_t^{(2)} & y_{t-2} > 3.25 \end{cases} \tag{6}$$

The opportunity to use a nonlinear times series model is difficult to establish. Indeed several authors as Tong (1990, pag. 217-221, pag. 362-375) and Fan and Yao (2003, pag.137-142) propose a preliminary analysis in order to detect nonlinearities in the data used. In order to present more of the functionalities and advantages to use **gretl**, in this section we show how to perform the same exploratory analysis. The following operations could be done using both the GUI and the **gretl** console, but for reason of space we only present the second use pattern. An early sign of nonlinear features is the presence of skewness and kurtosis, from Table 1 it is possible observe that the data present negative skewness (-0.36) and excess of kurtosis (-0.73).

Mean	Median	Minimum	Maximum	St. deviation	Skewness	Ex. kurtosis
6.6859	6.6475	3.6636	8.8524	1.2858	-0.36199	-0.73358

Table 1 Summary Statistics

A first set of operations as open the data set, transform the data, show them in a graph and present their main statistics could be done typing the following commands in the gretl console:

```
open http://ricardo.ecn.wfu.edu/pub/gretldata/lynx.gdt
series log_lynx=logs(lynx)
gnuplot log_lynx --time-series --with-lines --output=display
summary log_lynx
```

Another indication of nonlinearities is the presence of bimodality and nonormality in the data. This feature appears clearly from the histogram of the series and from the nonparametric estimation of the probability density function. In Figure 1, the left panel shows the histogram of the logged lynx data, whereas the right panel gives the kernel density estimation of the same data. From the histogram it is clear that the probability distribution is at least bimodal. Also the Doornik-Hansen chi-square test for normality is computed and the H_0 of normality is rejected at 1%. The kernel density estimation is a

[2] The data can be also downloaded from http://www.encyclopediaofmath.org/index.php/Canadian_lynx_data and then imported in gretl. The sample generally used is annual data from 1821 to 1919 with the variable transformed into log_{10}.

non-parametric way to estimate the probability density function of a random variable. If the estimated probability function of a variable does not approximate a normal distribution function it means the data are not normal. In our case we use the Gaussian Kernel density estimation but the Epanechnikov Kernel density estimation is also available.

Fig. 1 Histogram and Kernel density estimation

The lack of time-reversibility is another clue of nonlinearity. A way to study this feature is plotting the data in the conventional order as well as in the reversed time order. From left panel of Figure 2, it is clear that the lynx data are not time-reversible; there are asymmetric cycles, it took about six years to reach a peak from a trough and took only three or four years to drop from a peak to a trough. This is more circumstantial evidence that linear models as the ARMA models cannot be properly used on this time variable. The Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) only focus on the linear dependence existing between the lags of the variable studied. As it can be observed in right panel of Figure 2, the ACF and PACF show high autocorrelation in the data. Moreover the structure of the lags indicate the presence of a strong and persistent cycle in the data.

Fig. 2 Time plots of lynx data plus ACF and PACF

With linear ARMA model the only structure that could approximate this cycle is a AR(2) with two complex roots. This phenomena is called the pseudo-cycle and tends to disappear over the time. In case of the lynx data however the cycle is persistent. To perform Figure 1 and 2 in the **gretl** console, type:

```
freq log_lynx --normal --nbins=15   --show-plot
matrix d = kdensity(log_lynx)
gnuplot 2 1 --matrix=d --with-lines --output=display
genr log_lynx_reverted = sortby(-obs, log_lynx)
scatters log_lynx log_lynx_reverted --with-lines --output=display
corrgm log_lynx 40 --plot=display
```

Another feature that points to the use of nonlinear time series model is nonlinear dependence and the tool used is a scatter diagram of the variable with its lags. In the case of a bivariate normal distribution the scatter diagram should look like an ellipse with a decreasing density from the center of the cloud. All kinds of departure from this pattern indicate a nonormal bivariate distribution and so the presence of nonlinear dependence. In the Figure 3 a void is evident in the center of the diagram both at lags 1 and 2.

Fig. 3 Scatterplots of logged lynx and its lags from 1 to 2

In addition to the scatter diagrams some nonparametric estimation methods are employed to check nonlinear dependence. In **gretl** it is possible to use the locally weighted regression (loess) (Cottrell and Lucchetti, 2011, pags. 270-274) and The NadarayaWatson estimator (Cottrell and Lucchetti, 2011, pags. 271-274). Here we use the locally weighted regression (loess) method

to estimate a nonparametric regression of y_t on the lags y_{t-1}, y_{t-2}. In the regression of y_t on y_{t-2} two lines whose origin lies in the point cloud can be observed, that is a not normal behavior (and hence nonlinear), see Figure 4. This indicates a nonlinear dependence among y_t and its lagged value y_{t2}. Indeed in case of a bifurcation of the estimated nonparametric regression there is a nonlinear type relationship among variable and its lags. The following commands produce the scatters presented in the Figure 3 and 4.

```
lags 2; log_lynx
scatters log_lynx; log_lynx_1 log_lynx_2 --with-lines --output=display
gnuplot log_lynx log_lynx_1  --loess-fit  --output=display
gnuplot log_lynx log_lynx_2  --loess-fit  --output=display
```

Fig. 4 Nonparametric regression of logged lynx

4 How use the SETAR function of gretl

An easy way to install the SETAR function is to use the *function packages* window: $File > Function\,Files > On\,server$. Then from the list of all functions stored in the gretl's server install the SETAR function[3]. Once the function is installed, select it from $File > Function\,Files > On\,local\,machine...$, a dialogue box will appear and the user must specify some parameters needed for the execution of the function, as shown in Figure 5.

Fig. 5 Screenshot of the SETAR function call

The first five arguments are directly input into the estimation of the SETAR model defined in equation (6). Specifically, the box "Variable"is for choosing the dependent variable, the "Autoregressive Order" box allows the user to define the autoregressive order p and the "Delay" box allows the user to define the delay parameter d of threshold variable y_{t-d}. If no argument is introduced in the "Lags Included" box, all the lags from 1 to p of the SETAR model are included in the regression. Otherwise, the user can select the number of lags to include. If the user marks the box "Linearity Test", the $F(\hat{\gamma})$ test of equation (5) is executed using the bootstrap method proposed by Hansen (1996, 2000). If the option "Heteroscedasticity Correction" is selected, the SETAR model is estimated assuming the error term u_t is a heteroskedastic Martingale difference sequence. Once the dataset has been opened and the data transformed

[3] For more details on the use and edition of functions in **gretl** see Chapter 11 of Cottrell and Lucchetti (2011).

as described in the section 3, the user must reduce the sample up to 1919 to replicate Tong results[4]. In running the function, the results of the estimation appear in a new window. In this window, the results are divided into four main blocks: **i)** estimation of the linear AR(p), **ii)** main results of the SETAR model, **iii)** results for the superior regime of the model and **iv)** results for the inferior regime.

Output of the SETAR function[5]

```
**************************************************************
                 Threshold Autoregressive Model Estimate
                 OLS Standard Errors Reported
--------------------------------------------------------------
Dependent Variable  y                Sum Squared Errors  4.2623
Threshold Variable  y-2              Residual Variance   0.0468
Threshold Estimate  3.3101           Joint R-Squared     0.8659
0.95 Conf.Interval  [2.6117; 3.3589] AIC Inf.Criteria    -286.29
Heterosk_Test (pv)  0.4422           BIC Inf.Criteria    -275.42
**************************************************************
Regime1: y-2 <= 3.3101

            coefficient   std. error     z      p-value
  ---------------------------------------------------------
   const       0.618655    0.169645      3.647   0.0003   ***
   y-1         1.25815     0.0729194    17.25    1.05e-66 ***
   y-2        -0.434315    0.0886438    -4.900   9.61e-07 ***

   Observations = 66
   Freedom Degrees = 63
   Sum Squared Errors = 2.54956
   Residual Variance = 0.0404692
   F-test p-value = 7.57954e-30
   R-squared = 0.881001
   AIC Inf.Criteria = -205.676
   BIC Inf.Criteria = -199.107

0.95 Confidence Regions for Parameters:

Variable             Low              High
-----------------------------------------------
const               -0.06045      1.17000
y-1                  1.07504      1.62569
```

[4] The estimation of the SETAR model may be completed with the gretl *command line interface* simply typing this command: smpl 1821 1919 namelist = SETAR(log_lynx, 2, 2, null, 0, 0), where *log_lynx* is the name of the dependent variable.

[5] Here for reasons of space the first block of the output is skipped.

```
y-2                -0.89958    -0.11708
*****************************************************************

Regime1: y-2 > 3.3101

            coefficient   std. error     z      p-value
  ---------------------------------------------------------
   const       1.14693      0.964777     1.189   0.2345
   y-1         1.59192      0.119682    13.30    2.28e-40 ***
   y-2        -1.00004      0.293768    -3.404   0.0007   ***

   Observations = 31
   Freedom Degrees = 28
   Sum Squared Errors = 1.71272
   Residual Variance = 0.0611687
   F-test p-value = 1.57078e-11
   R-squared = 0.830841
   AIC Inf.Criteria = -80.6177
   BIC Inf.Criteria = -76.3157

0.95 Confidence Regions for Parameters:

Variable             Low              High
-----------------------------------------
const               -1.80364     3.66965
y-1                  1.22498     1.90369
y-2                 -1.78327    -0.14210
Generated list namelist
*****************************************************************
```

The estimation by LS of the linear AR(p) is presented because it is the initial part of the methodology to estimate the SETAR model. Thus, in this first block the estimated coefficients, the standard deviations, the t-ratios (z) and the p-values are shown together with some general statistics of the regression. In the second block, some parameters of the whole SETAR model are shown, such as the threshold parameter, $\hat{\gamma}$, and its confidence interval $\hat{\Gamma}$. Here, to help the user decide whether to use a heteroscedasticity correction in the regression, the Heteroskedasticity Test of White is also shown. In the third and fourth blocks each regime of the SETAR model estimated is presented joint to respective confidence interval $\hat{\Theta}_{\phi}$. In this example, using the Lynx data, the estimated autoregressive coefficients are very close to those in equation (6), and the estimated threshold parameter, $\hat{\gamma}$, differs only in decimal places from Tong's estimation. To guarantee reliability of our estimation, we have compared them with those obtained in Matlab and Gauss, with the orig-

inal code of Hansen[6]. As expected, we obtain exactly the same results. It is interesting to highlight that we have a parametric estimation of γ compatible with that of Tong, given that this value is inside our $\hat{\Gamma}$ confidence interval $[2.6117; 3.3589]$. Therefore, we have more precision and additional information about this parameter.

The information in this output window give the user the possibility to make some inference on the estimated model but the main tool needed when choosing a particular specification for the SETAR model remains the **linearity test**. The $F(\hat{\gamma})$ test of the equation (5), can be performed easily by marking the box "Linearity Test" in the function call box once all the parameters of the SETAR model have been inserted. When the Heteroskedasticity Correction is selected, then the **linearity test** is performed assuming heteroscedasticity in the error term. In both cases, the results of the test are showed at the bottom of the main output window. In the output the statistic of the $F(\hat{\gamma})$ test plus the p-value calculated by bootstrap is reported[7].

```
**************************************************************
   Test of Null of No Threshold Against Alternative of Threshold
    Under Maintained Assumption of Homoskedastic Errors
**************************************************************
Number of Bootstrap Replications 1000
Trimming Percentage 0.15
Threshold Estimate 3.3101
LM-test for no threshold 27.846
Bootstrap P-Value 0.001
```

5 Conclusions

This paper demonstrated how to work with the SETAR models in **gretl**. Using *hansl*, its program language, we developed a function that can be easily used to estimate a SETAR model and make diagnostics. The starting point is the procedure to estimate a SETAR model with two regimes, originally proposed by Hansen (1997). Thus, we implemented all of the Hansen methodology in our function and we presented how use the several features of gretl to perform an exploratory analysis of the data.

Using free and open source software has many advantages for the academic and scientific world. First, due to the free availability of the code, users can study the code to see how it works, allowing users to detect faults and errors in it and correct these mistakes quickly for a more efficient the software. Second, transparency and the right to modify the software encourage community participation and inclusion in the project and the use of the program. Users

[6] The original Matlab (or Gauss) code of Hansen (2000) can be downloaded directly from the homepage of this author: http://www.ssc.wisc.edu/~bhansen/progs/progs_threshold.html.

[7] In case the user prefers to work on the CLI he only needs to add the following command: namelist = SETAR(log_ynx, 2, 2, null, 0, 1) to execute the **linearity test**.

can easily request further extensions or features for the software or participate actively in the development process. Third, the possibility to use an easy to use GUI to estimate a SETAR model should boost its diffusion in academic and higher education institutions. Our use of the Lynx data, from the exploratory analysis up to the estimation and diagnostics, could be employed as short guide for teaching purposes.

Acknowledgements The first author gratefully acknowledges financial support from Regione Autonoma della Sardegna through *Programma Master and Back*. Financial support from research project ECO2010-15332 from Ministerio de Ciencia e Innovacion, and from UPV/EHU *Econometrics Research Group*, Basque Government grant IT-642-13, is gratefully acknowledged by the second author.

References

L. Adkins. Using gretl for principles of econometrics, 3rd edition version 1.313, 2010.

G. Baiocchi and W. Distaso. Gretl: Econometric software for the gnu generation. *Journal of applied econometrics*, 18:105–110, 2003.

K. S. Chan. Consistency and limiting distribution of a least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, 21:520533, 1993.

A. Cottrell and R. Lucchetti. **gretl** users guide gnu regression, econometrics and time-series library, 2011.

Y. Fan and Q. Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. SPRINGER, New York, 2003.

B. Hansen. Inference when a nuisance parameter is not identified under the null hypotesis. *Econometrica*, 64:413–430, 1996.

B. Hansen. Inference in tar models. *Studies in Nonlinear Dynamics and Econometrics*, 1:895–904, 1997.

B. Hansen. Sample splitting and threshold estimation. *Econometrica*, 2000.

R. Smith and J. Mixon. Teaching undergraduate econometrics with gretl. *Journal of Applied Econometrics*, 21(7), 2006.

H. Tong. *Threshold models in non-linear time series analysis*. Lecture Notes in Statistics, No.21. Springer, Heidelberg, 1983.

H. Tong. *Nonlinear time series, a dynamical system approach*. Oxford University Press, London, 1990.

H. Tong and K. Lim. Threshold autoregression, limit cycles and cyclical data (with discussion). *Journal of the Royal Statistical Society*, B42:245–292, 1980.

A. Yalta and Y. Yalta. Gretl 1.6.0 and its numerical accuracy. *Journal of applied econometrics*, 22:849–854, 2007.