INSTRUCTIONS TO THE PROGRAM OF ESTIMATING TAR

Developed by **Konstantin Gluschenko**
(the very first version of the program was written by Peter Rostovtsev)
glu@nsu.ru
http://econom.nsu.ru/users/gluschenko
Version: 09.2005

0. Introduction

This program estimates **threshold autoregressions** (TAR), using the Microsoft Excel environment. The program is organized as an Excel macro titled *TARmodel*. The program is written in VBA (Visual Basic for Applications).

To see or modify the source code of the program, use the Excel main menu: Tools → Macro → Macros → Edit.

1. Algorithm

The model under consideration is

$$\Delta y_t = \begin{cases} \lambda_{(out)}(y_{t-1} - c_{(+)}) + \varepsilon_{(out)t} & \text{if } y_{t-1} > c_{(+)} \\ \lambda_{(in)} y_{t-1} + \varepsilon_{(in)t} & \text{if } c_{(+)} \geq y_{t-1} \geq c_{(-)} \\ \lambda_{(out)}(y_{t-1} - c_{(-)}) + \varepsilon_{(out)t} & \text{if } y_{t-1} < c_{(-)} \end{cases} \qquad (t = 1, \ldots, T),$$

$$-2 < \lambda_{(out)} < 0$$

where $\{y_t\}_{t=0,\ldots,T}$ is a time series, Δ is the first difference operator ($\Delta y_t \equiv y_t - y_{t-1}$), and $\varepsilon_{(.)t}$ are regression residuals. Parameters to be estimated are $\lambda_{(out)}$ and $\lambda_{(in)}$, the autoregression coefficients (note that the estimate of $\lambda_{(in)}$ is generally of no interest), and $c_{(+)}$ ans $c_{(-)}$, the upper and lower threshold, respectively. Restriction $\lambda_{(in)} = 0$ can be imposed, implying the process within the band $[c_{(-)}, c_{(+)}]$ to be a pure random walk.

Thresholds c are estimated subject to $W \leq t_{(out)}/(T+1) \leq 1 - W$ (similarly for $t_{(in)}$), where W is a (user-)specified window, $W \leq 0.5$, $t_{(out)}$ is the number of observations such that $y_t \notin [c_{(-)}, c_{(+)}]$, and $t_{(in)}$ relates to $y_t \in [c_{(-)}, c_{(+)}]$; $t_{(out)} + t_{(in)} = T+1$. Following Andrews (1993), the window would be not less than 0.15, that is, both parts of the series contain not less than 15% of observations. However, the decision as to the value of W should be made depending on the size of the analyzed series. (Having short series, it is reasonable to widen the window; long series allow decreasing W.)

For discussion of economic and econometric meaning of this model, see the literature, e.g., Obstfeld and Taylor (1997a, b), Gluschenko (2004), etc. (Commonly, y_t is a price/price level differential between two locations, i.e., $y_t = \ln(p_{rt}/p_{st})$, where p_{it} is a price/price level in location i.)

Testing for the threshold effect is, in fact, a specification test of TAR vs. AR. That is, the hypothesis to be tested is H_0: the data generating process is AR(1) with parameter λ_0, $\Delta y_t = \lambda_0 y_{t-1} + \varepsilon_t$ ($t = 1, \ldots, T$), against the alternative H_a: the process is TAR with parameters $\lambda_{(out)}$, $\lambda_{(in)}$, and c.

The program uses the estimation and testing method put forward by Obstfeld and Taylor (1997a, b) in their Appendix A, slightly modifying it. Three modifications are made: (*a*) Instead of fixed W =

0.1, the window can have an arbitrary value from [0, 0.5]. (*b*) The step of the grid search is the difference between neighboring values of the observations rather than any fixed value (e.g., 0.001). (*c*) While the common practice is the use of symmetric thresholds, $c_{(-)} = -c_{(+)}$, these are asymmetric, though related by the relationship $c_{(-)} = \ln(2 - \exp(c_{(+)}))$; for motivation, see Gluschenko (2004), Appendix A. (It worth noting that y_t is deemed to be logarithm.)

To estimate and test TAR, a best-fit grid-search on the threshold parameter c is used. The objective function is the log ratio of the estimated likelihood function of the TAR and that of the AR(1), $LLR = L_{TAR} - L_{AR}$, which is maximized. In turn,

$$L_{AR} = L(\lambda_0, \sigma) = -\frac{1}{2}\sum_{t=1}^{T}(\ln 2\pi + \ln \sigma^2 + \varepsilon_t^2 / \sigma^2),$$

$$L_{TAR} = L(\lambda_{(out)}, \lambda_{(in)}, \sigma_{(out)}, \sigma_{(in)}, c) = -\frac{1}{2}\sum_{t\in O}(\ln 2\pi + \ln \sigma_{(out)}^2 + \varepsilon_{(out)t}^2 / \sigma_{(out)}^2) -$$

$$-\frac{1}{2}\sum_{t\in I}(\ln 2\pi + \ln \sigma_{(in)}^2 + \varepsilon_{(in)t}^2 / \sigma_{(in)}^2),$$

where $O = \{t: y_{t-1} \notin [c_{(-)}, c_{(+)}]\}$, and $I = \{t: y_{t-1} \in [c_{(-)}, c_{(+)}]\}$.

The procedure of estimating is as follows.

1. Estimate AR(1) by OLS, obtaining $\hat{\lambda}_0$ and $\hat{\sigma}$, and calculating $\hat{L}_{AR}$.

2. For each t compute $c_t = y_t$ if $y_t > 0$, or $c_t = \ln(2 - \exp(y_t))$ if $y_t < 0$. Sort $\{c_t\}$ in ascending order, obtaining $\{c_k\}$. Compute $k0 = [W(T - 1)] + 1$ and $k1 = T - k0$. (Recall that W is the window; $[x]$ means here the integer part of x.)

3. For each $k = k0,\ldots,k1$: compute $c_{(+)k} = c_k$ and $c_{(-)k} = \ln(2 - \exp(c_k))$; construct sets $O_k = \{t: y_{t-1} \notin [c_{(-)k}, c_{(+)k}]\}$ and $I_k = \{t: y_{t-1} \in [c_{(-)k}, c_{(+)k}]\}$; estimate $\Delta y_t = \lambda_{(out)k} y_{t-1} + \varepsilon_{(out)t}$ $(t \in O_k)$ and $\Delta y_t = \lambda_{(in)k} y_{t-1} + \varepsilon_{(in)t}$ $(t \in I_k)$ by OLS; compute LLR_k.

4. Take $k^* = \text{argmax}_k(LLR_k)$ and $LLR^* = LLR_{k^*}$. Take $\hat{\lambda}_{(out)} = \lambda_{(out)k^*}$, $\hat{\lambda}_{(in)} = \lambda_{(in)k^*}$, and $\hat{c} = c_{k^*}$.

5. Estimate p-value of LLR, using a model-based bootstrap to obtain the distribution of the LLR statistic under the null hypothesis. The procedure is as follows. In each replication i ($i = 1,\ldots,N$), $T + 50$ random values are generated, $\varepsilon_{-50}^{(i)},\ldots,\varepsilon_T^{(i)}$, $\varepsilon_t^{(i)} \sim i.i.d.N(0, \hat{\sigma}^2)$. Then a simulated time series is generated as $y_t^{(i)} = (\hat{\lambda} + 1)y_{t-1}^{(i)} + \varepsilon_t^{(i)}$, $t = -50,\ldots,T$ and $y_{-51}^{(i)} = 0$, discarding the first 50 "observations" to avoid an initial value bias. The AR and TAR models are estimated over this time series like in steps 1 through 4, so yielding a realization of $LLR^{(i)}$. The number of realizations such that $LLR^{(i)} > LLR^*$ related to the number of replications, N, is the p-value of the LLR^*, i.e., the probability to accidentally obtain a value of LLR greater than LLR^*, provided that the null hypothesis holds.

2. The input data

The program uses the input data from an Excel sheet. It need not be in the same Excel file that contains the program; the only condition is that the latter should be opened in parallel with the file containing the data to be processed.

A time series should be organized as a column vector; that is, a single time series is set to a column of an Excel sheet, the number of rows equaling the sample size (T+1), and the order of rows

corresponding the order of periods, *t*. The program can process an arbitrary set of time series with the same sample size in one run. The series in the set can be located horizontally (in "series columns") and/or vertically (in "tiers"). The series should not be adjacent; the program skips empty series ("gaps"). However, vertical gaps (empty tiers) should have the same length as the actual series (i.e., the number of rows in an empty tier should equal the sample size).

The series should not occupy the first row (row 1) and the first column (column A); identifiers (names, codes, etc.) are placed in these row and column. Row 1 contains identifiers of series columns; and column A contains identifiers of tiers in the first row of each tier. The beginning of the data range (the rectangle of an Excel sheet that contains series to be processed) can be shifted from the first row of the sheet by an arbitrary number of rows; the same is valid for columns.

The reason for such an organization of data is that the program has been originally developed to process sets of pairwise price differentials across locations. Figure 1 demonstrates an example of organization of such a data.

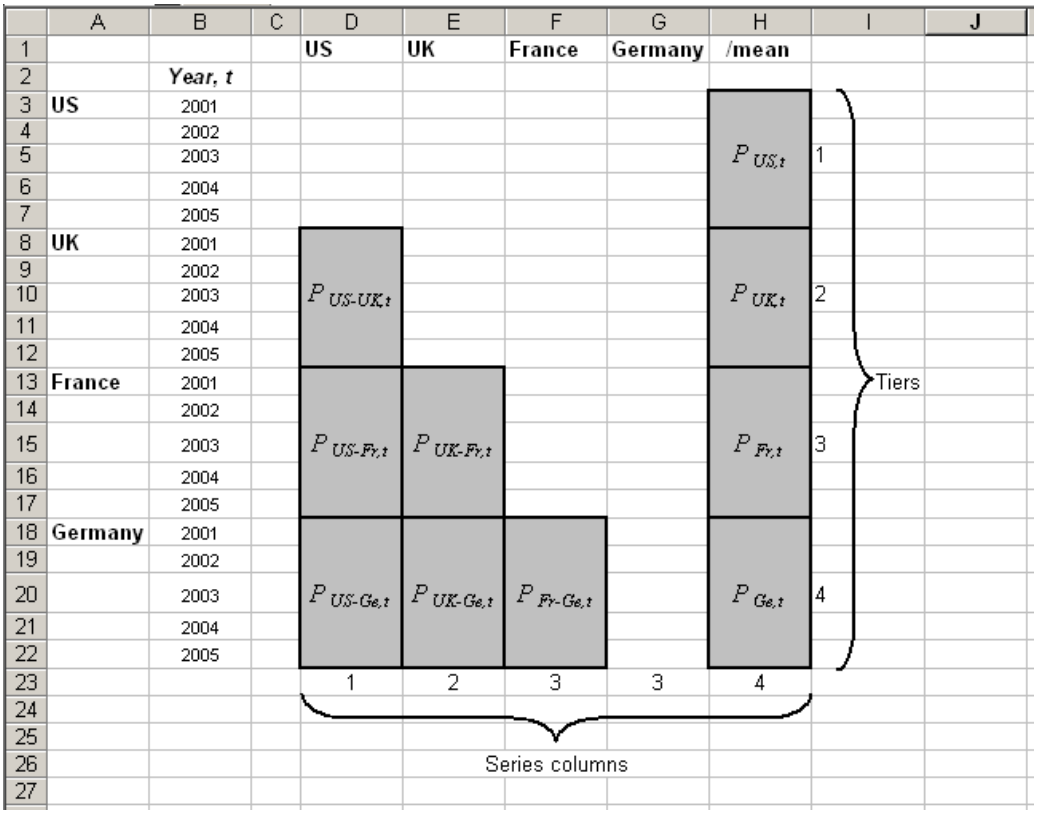

Figure 1. An example of organization of data (each gray rectangle represents a time series).

Any subset of series, which falls into a rectangle range, can be chosen to be one-run processed. For example, four series of prices related to the cross-country mean will be analyzed if the range H3:H22 is chosen; provided that the range is D8:F22, six pairwise series will be processed. All displayed series are processed with the range D3:H22; the only series $\{P_{UK\text{-}Ge,t}\}$ will be processed, given the range of E18:E22.

It is possible to estimate TAR over a part of a series rather than over the whole series. To do so, it needs to choose a range covering the processed part of the series. For example, specifying the range H9:H12, the estimation involves $\{P_{UK,t}\}$over 2002 to 2005; the range H8:H11 yields the estimation

without recent data; and the range H9:H11 provides estimates over the middle part of $\{P_{UK,t}\}$, excluding 2001 and 2005. Such estimations can be performed across a number of series in one run, but these obviously can be only series from the same tier. For example, the range D14:H17 specifies estimation for three series $\{P_{US\text{-}Fr,t}\}$, $\{P_{UK\text{-}Fr,t}\}$, and $\{P_{Fr,t}\}$ without the initial year. However, to obtain estimates for all four series involving France, one more run is needed: that with the range F19:F22.

To specify series to be processed, a relevant series or a set of series can be highlighted before starting the program. Figure 2 provides an example. In this example, the range D3:I17 is highlighted. Then six series will be processed; omitted are series that involve Germany.

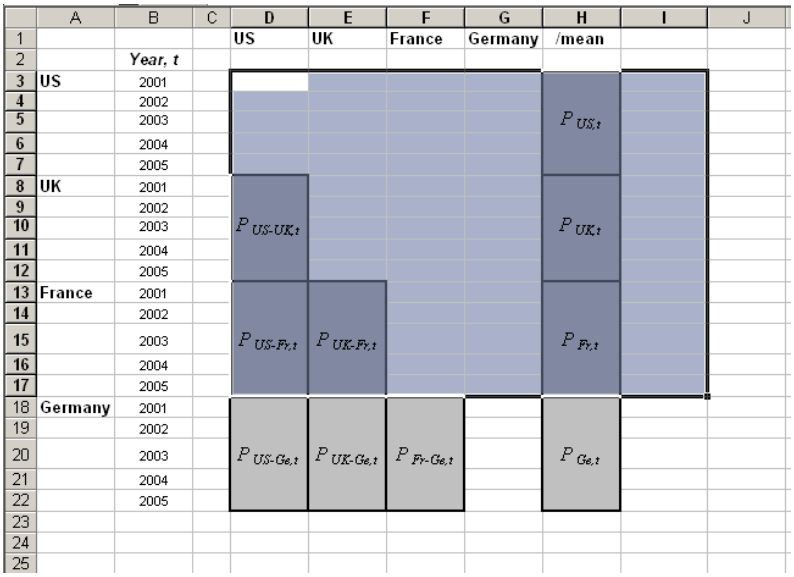

Figure 2. Specifying time series to be processed in one run.

An alternative way is to specify an Excel sheet range (that contains the data to be processed) in the dialog box of the program; see below.

3. Starting the program

With the use of the Excel main menu, the program is called by the sequence Tools → Macro → Macros → Run. (If other macros are active, highlight TARmodel before clicking Run). An alternative way is to press <Ctrl>t. After that the dialog box of the program appears, as displayed in Figure 3.

At its appearance, the dialog box contains default settings. You may change all or some of the settings. The elements of the dialog box are as follows.

Input range specifies the Excel sheet range that contains series to be processed. By default, this is the range that has been being highlighted before starting the program. If there is no a highlighted range, the address of the active sell appears. Then you should manually input coordinates of the series range. You may also specify a different range instead of the highlighted one.

Sample size specifies the length of a single series (T+1). By default, it is equal to the number of rows in the highlighted range. (It there is no such a range, the value in this box is 1.) If the range is manually specified or changed, the actual sample size should be input either directly or with the use of the spin buttons at the right of this box. The default sample size should also be changed

to the actual value when there is more than one tier in the input range. (In the example in Figure 3, the sample size should be changed to 5.)

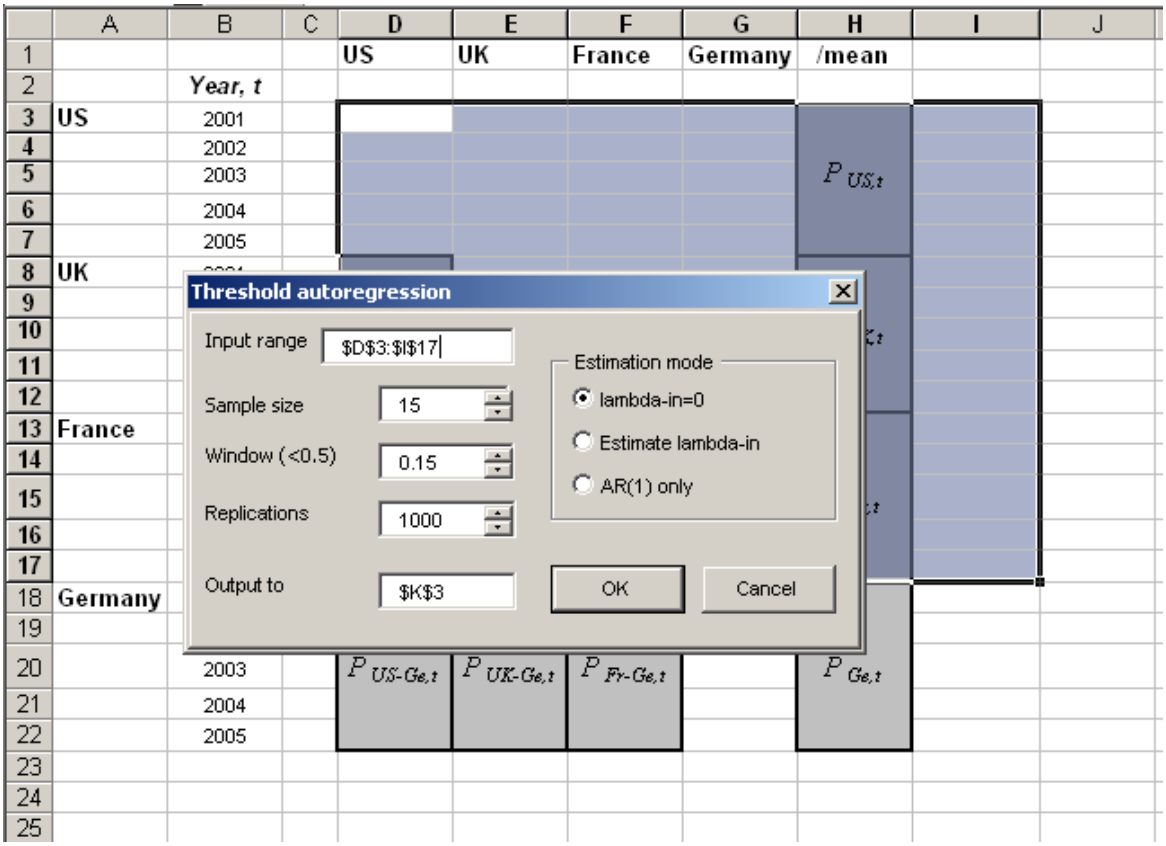

Figure 3. Dialog box of the program.

Window specify the value of W (see section 1), that is, the minimal portion of the series that may be beyond/within the band $[c_{(-)}, c_{(+)}]$. The default value is 0.15.

Replications specifies the number of Monte Carlo experiments, N, while estimating the p-value of LLR. The default value is 1,000. The p-value is rather sensitive to N; and so, it is desirable to have N as large as possible. But on the other hand, it is the value of N that determinates the program run time. Thus, the choice would depend on the number of processed series. Anyway, N=1,000 seems not enough; I would recommend at least 10,000.

Output to specifies the upper left cell of the output table. By default, the output table starts from the same row as the processed series; its starting column is shifted by one column to the right from the end of the input range. You may specify any different location, provided that the output range does not overlap the input range.

Estimation mode contains three option buttons. The default option is lambda-in=0, implying that the estimations will be performed subject to restriction $\lambda_{(in)} = 0$. This restriction is removed with the option Estimate lambda-in. At last, only AR(1) model is estimated instead of both AR(1) and TAR, if the option AR(1) only is chosen.

The program checks the accuracy of settings while inputting them. When the setting in the dialog box are proper, click OK. Then the dialog box becomes elongated, as Figure 4 demonstrates. Its additional part displays the structure of the input range, and provides a possibility to correct the settings.

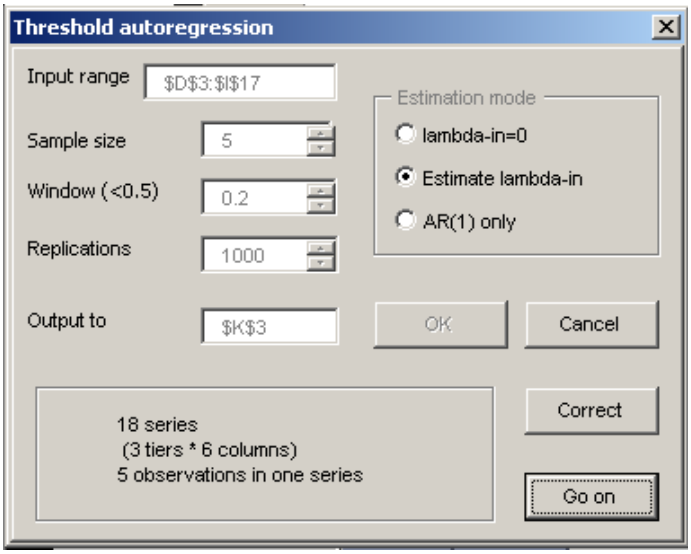

Figure 4. Expanded dialog box.

Figure 4 uses the same example like Figure 3. As seen, the value of *W* (window) is changed as well as the sample size. Now the program understands that series are located both one after another and one below another, suggesting that there are 18 series in the input range (including empty ones that will be skipped while running).

You may return to specifying the settings, clicking Correct; or you may start estimating, clicking Go on.

While the program is working, you can exit at any time by pressing `Esc`.

4. Output

Provided that a set of time series is specified, the series are processed from left to right and top-down, empty series being skipped. An example of the output table is presented in Figure 5.

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
1																							
2																							
3	N exp	r	s	p(unit root)	AR lambda	s.d.	t-stat	AR T-half	LLR	p(LLR)	TAR lambda-out	s.d.	t-stat	TAR T-half	c(-1)	C(-1), %	c	C, %	t out	t in	TAR lambda-in	s.d.	t-stat
4	10000	US	/mean		-0.69838378	0.1	-6.675	0.578297	20.73618	0.001252	-1.901853444	0.402	-0.24	6.70982926	0.06	5.698961	0.06	6.07	18	65	-0.286290883	0.146	-1.96
5	10000	UK	US		-0.46895766	0.09	-5.081	1.095169	12.57938	0.035178	-0.636832009	0.14	-4.54	0.68432637	0.02	1.682556	0.02	1.769	58	25	-1.715142755	0.541	-3.17
6	10000	UK	/mean		-0.40063251	0.09	-4.575	1.354119	20.04193	0.001632	-1.429829883	0.418	-1.36	0.82090867	0.08	8.411959	0.08	8.491	13	70	-0.316614984	0.119	-2.66
7	10000	France	US		-0.37630329	0.09	-4.316	1.468249	6.939896	0.334304	-0.825088301	0.146	-5.63	0.39756668	0.04	4.567948	0.05	4.665	38	45	-0.106563386	0.31	-0.34
8	10000	France	UK		-0.27572417	0.08	-3.665	2.148741	58.28129	0	-0.798027834	0.238	-3.36	0.43331844	0.1	10.78822	0.1	10.8	33	50	-0.023347652	0.099	-0.23
9	10000	France	/mean		-0.01417772	0.02	-0.922	48.54249	8.377187	0.248852	-0.212414925	0.116	-1.83	2.9028223	0.25	28.73156	0.25	28.76	20	63	-0.007338084	0.022	-0.34
10	10000	Germany	US		-0.01048411	0.01	-0.734	65.76688	9.15148	0.200922	-0.119351616	0.055	-2.16	5.45369316	0.18	20.03098	0.18	20.13	54	29	0.050059591	0.037	1.367
11	10000	Germany	UK		-0.00891599	0.01	-1.067	77.39498	29.37428	0.052935	-0.092295148	0.127	-0.73	7.15794891	0.43	53.57289	0.43	53.66	12	71	-0.003407009	0.007	-0.47
12	10000	Germany	France		-0.0060403	0.01	-0.526	114.4069	13.20381	0.186821	-0.460226178	0.155	-2.9	1.15862788	0.4	49.86947	0.41	49.98	27	56	0.003715117	0.017	0.216
13	10000	Germany	/mean		-0.00303045	0.02	-0.193	228.3808	7.684764	0.34926	-0.276259299	0.135	-2.04	2.14382865	0.19	20.57084	0.19	20.74	22	61	0.03673584	0.021	1.77
14																							

Figure 5. Example of the output table.

The items in the output table are as follows.

N exp displays a current number of the Monte Carlo experiment, *i*, while processing a given series; after completing the estimation, this column contains the total number of replications, *N*.

r and **s** are identifiers of the series column and tier. For example, if all the 10 series from Figure 1 would be processed in one run (specifying the input range as D3:H22), these columns of the output table look like in Figure 5.

p(unit root) is always empty. This column is destined for *p*-values of a unit root test, the values being obtained by user with the use of any other program (e.g., MacKinnon's (1996) program, EViews, etc.).

AR lambda is the estimate of λ_0.

s.d. is the standard deviation of λ_0.

t-stat is the value of *t*-statistic for the estimate of λ_0.

AR T-half means the half-life in the AR(1) model; it is computed as $\ln 0.5/\ln(|1+\lambda_0|)$; its unit of measure coincides with the frequency of the corresponding time series (months, quarters, etc.).

LLR is the estimated value of the LLR statistic for TAR.

p(LLR) is the *p*-value of the estimated LLR.

TAR lambda-out is the estimate of $\lambda_{(out)}$.

s.d. is the standard deviation of $\lambda_{(out)}$.

t-stat is the value of *t*-statistic for the estimate of $\lambda_{(out)}$.

TAR T-half means the half-life in the TAR model; it is computed as $\ln 0.5/\ln(|1+\lambda_{(out)}|)$.

c(–1) is a value of c that is the nearest to the estimate of c. Since the threshold grid is discrete, we do not know the exact point of the maximum of *LLR*. Thus, a true value of c lies somewhere between the estimate of c and the neighboring nod of the grid. The value in this column gives an idea of how large can be inaccuracy in the estimate of c.

C(–1), % is the percentage of $c(-1)$; it is calculated as $(e^{c(-1)} - 1)\cdot 100$. (Recall that $\{y_t\}$ are deemed to be logarithms.)

c is the estimate of c (more exactly, of $c_{(+)}$).

C, % is the percentage of c; it is calculated as $(e^{c} - 1)\cdot 100$.

t out is $t_{(out)}$, that is, the number of observations such that $y_t \notin [c_{(-)}, c_{(+)}]$.

t in is $t_{(in)}$, that is, the number of observations such that $y_t \in [c_{(-)}, c_{(+)}]$.

TAR lambda-in is the estimate of $\lambda_{(in)}$; if the option is lambda-in=0, the value in this column is always 0.

s.d. is the standard deviation of $\lambda_{(in)}$.

t-stat is a value of *t*-statistic for the estimate of $\lambda_{(in)}$; if the option is lambda-in=0, the value in this column is always 0.

References

Andrews, D. W. K. (1998). Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica*, **61** (4), 821-856.

Gluschenko, K. (2004). The Law of One Price in The Russian Economy. *LICOS Discussion Paper* No. 152/2004. (Available on www.econ.kuleuven.ac.be/licos/DP/DP2004/DP152.pdf)

MacKinnon, J. G. (1996). Numerical Distribution Functions for Unit Root and Cointegration Tests. *Journal of Applied Econometrics*, **11** (6), 601-618.

Obstfeld, M., and A. M. Taylor (1997a). Non-Linear Aspects of Good-Market Arbitrage and Adjustment: Heckscher's Commodity Points Revisited. *CEPR Discussion Paper* No. 1672.

Obstfeld, M., and A. M. Taylor (1997b). Non-Linear Aspects of Good-Market Arbitrage and Adjustment: Heckscher's Commodity Points Revisited. *Journal of Japanese and International Economies*, **11**, 441-479.