Title

stata.com

threshold — Threshold regression

Description

`threshold` extends linear regression to allow coefficients to differ across regions. Those regions are identified by a threshold variable being above or below a threshold value. The model may have multiple thresholds, and you can either specify a known number of thresholds or let `threshold` find that number for you through the Bayesian information criterion (BIC), Akaike information criterion (AIC), or Hannan–Quinn information criterion (HQIC).

Quick start

Threshold regression model for the dependent variable `y` with region-dependent intercepts for two regions of `x`

```
threshold y, threshvar(x)
```

Add the first lag of `x` as a region-invariant variable

```
threshold y l.x, threshvar(x)
```

Add the first lag of `y` as a region-dependent variable

```
threshold y l.x, threshvar(x) regionvars(l.y)
```

Threshold regression model of `y` with region-dependent intercepts for three regions determined by two threshold values of `x`

```
threshold y, threshvar(x) nthresholds(2)
```

Use BIC to select the number of thresholds from a maximum of 5 thresholds

```
threshold y, threshvar(x) optthresh(5)
```

Menu

Statistics > Time series > Threshold regression model

Syntax

```
threshold depvar [indepvars] [if] [in], threshvar(varname) [options]
```

indepvars is a list of variables with region-invariant coefficients.

options	Description
Model	
* `threshvar(`*varname*`)`	threshold variable
`regionvars(`*varlist*`)`	include region-varying coefficients for specified covariates
`consinvariant`	replace region-varying constant with a region-invariant constant
`noconstant`	suppress region-varying constant terms
`trim(#)`	trimming percentage; default is `trim(10)`
`nthresholds(#)`	number of thresholds; default is `nthresholds(1)`; not allowed with `optthresh()`
`optthresh(#[ , `*ictype*`])`	select optimal number of thresholds less than or equal to #; not allowed with `nthresholds()`
SE/Robust	
`vce(`*vcetype*`)`	*vcetype* may be `oim` or `robust`
Reporting	
`level(#)`	set confidence level; default is `level(95)`
`nocnsreport`	do not display constraints
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
`nodots`	suppress replication dots
`dots(#)`	display dots every # replications
Advanced	
`ssrs(`*stub*`*` | *newvarlist*`)`	create variable with sum of squared residuals (SSRs) for each tentative threshold
`constraints(`*numlist*`)`	apply specified linear constraints; not allowed with `optthresh()`
`coeflegend`	display legend instead of statistics

* `threshvar()` is required.

You must `tsset` your data before using `threshold`; see [TS] **tsset**.

indepvars and *varlist* may contain factor variables; see [U] **11.4.3 Factor variables**.

depvar, *indepvars*, *varlist*, and *varname* may contain time-series operators; see [U] **11.4.4 Time-series varlists**.

`by`, `rolling`, and `statsby` are allowed; see [U] **11.1.10 Prefix commands**.

`coeflegend` does not appear in the dialog box.

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

ictype	Description
`bic`	Bayesian information criterion (BIC); the default
`aic`	Akaike information criterion (AIC)
`hqic`	Hannan–Quinn information criterion (HQIC)

Options

Model

`threshvar(`*varname*`)` specifies the variable from which values are to be selected as thresholds. `threshvar()` is required.

`regionvars(`*varlist*`)` specifies additional variables whose coefficients vary over the regions defined by the estimated thresholds. By default, only the constant term varies over regions.

`consinvariant` specifies that the constant term should be region invariant instead of region varying.

`noconstant` suppresses the region-varying constant terms (intercepts) in the model.

`trim(#)` specifies that `threshold` treat the value at the #th percentile of the threshold variable as the first possible threshold and the value at the $(100-\#)$th percentile as the last possible threshold. # must be an integer between 1 and 49. The default is `trim(10)`.

`nthresholds(#)` specifies the number of thresholds. Specifying the number of thresholds is equivalent to specifying the number of regions because the number of regions is equal to $\#+1$ thresholds. The default is `nthresholds(1)`, equivalent to 2 regions.

`optthresh(#` [`,` *ictype*]`)` specifies that `threshold` choose the optimal number of thresholds, up to a possible #. By default, the optimal number of thresholds is based on the BIC, but you may specify the information criterion (*ictype*) to be used. *ictype* may be `bic` (the default), `aic`, or `hqic`.

SE/Robust

`vce(`*vcetype*`)` specifies the type of standard error reported, which includes types that are derived from asymptotic theory (`oim`) and that are robust to some kinds of misspecification (`robust`); see [R] ***vce_option***.

Reporting

`level(#)`, `nocnsreport`; see [R] **estimation options**.

display_options: `noci`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(#)`, `fvwrapon(`*style*`)`, `cformat(%`*fmt*`)`, `pformat(%`*fmt*`)`, `sformat(%`*fmt*`)`, and `nolstretch`; see [R] **estimation options**.

`nodots` suppresses display of the replication dots. By default, one dot character is displayed for each successful replication. A red 'x' is displayed if *command* returns an error.

`dots(#)` displays dots every # replications. `dots(0)` is a synonym for `nodots`.

Advanced

`ssrs(`*stub*`*`|*newvarlist*`)` creates a variable containing the sum of squared residuals (SSRs) that was computed for each tentative threshold value during the search for the kth threshold. For observations where the value of the threshold variable specified in `threshvar()` is not a tentative threshold, the corresponding value of the variable created by `ssrs()` for that observation will be missing.

If you specify *stub*`*`, Stata will create k new variables with the names *stub*`1`, ..., *stubk*, which will contain the SSRs for the 1st, ..., kth thresholds, where k is the # specified in `nthresholds()` or the optimal number of thresholds if `optthresh()` is specified.

If you specify a list of new variable names, you may request SSRs for up to the # specified in `nthresholds()`. If you specify `optthresh(#)` and the optimal number of thresholds is less than #, any additional variables will contain only missing values.

`constraints(`*numlist*`)` specifies the constraints by number after they have been defined by using the `constraint` command; see [R] **constraint**. `constraints()` may not be specified with `optthresh()`

The following option is available with `threshold` but is not shown in the dialog box:

`coeflegend`; see [R] **estimation options**.

Remarks and examples

stata.com

`threshold` extends linear regression to allow coefficients to differ across regions. Those regions are identified by a threshold variable being above or below a threshold value. The model may have multiple thresholds, and you can either specify a known number of thresholds or let `threshold` find that number for you by minimizing an information criterion.

These models are good alternatives to linear models for capturing abrupt breaks or asymmetries observed in most macroeconomic time series over the course of a business cycle. Common threshold regression models include the threshold autoregression model and self-exciting threshold model. In the threshold autoregression model, proposed by Tong (1983), the dependent variable is a function of its own lags; see Tong (1990) for details. In the self-exciting threshold model, the lagged dependent variable is used as the threshold variable. For a survey of threshold regression models in economics, refer to Hansen (2011).

Formally, consider a threshold regression with two regions defined by a threshold γ. This is written as

$$y_t = \mathbf{x}_t\boldsymbol{\beta} + \mathbf{z}_t\boldsymbol{\delta}_1 + \epsilon_t \qquad \text{if} \quad -\infty < w_t \leq \gamma$$
$$y_t = \mathbf{x}_t\boldsymbol{\beta} + \mathbf{z}_t\boldsymbol{\delta}_2 + \epsilon_t \qquad \text{if} \quad \gamma < w_t < \infty$$

where y_t is the dependent variable, $\mathbf{x}_t$ is a $1 \times k$ vector of covariates possibly containing lagged values of y_t, $\boldsymbol{\beta}$ is a $k \times 1$ vector of region-invariant parameters, ϵ_t is an IID error with mean 0 and variance σ^2, $\mathbf{z}_t$ is a vector of exogenous variables with region-specific coefficient vectors $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$, and w_t is a threshold variable that may also be one of the variables in $\mathbf{x}_t$ or $\mathbf{z}_t$.

The parameters of interest are $\boldsymbol{\beta}$, $\boldsymbol{\delta}_1$, and $\boldsymbol{\delta}_2$. Region 1 is defined as the subset of observations in which the value of w_t is less than the threshold γ. Similarly, Region 2 is defined as the subset of observations in which the value of w_t is greater than γ. Inference on the nuisance parameter γ is complicated because of its nonstandard asymptotic distribution; see Hansen (1997, 2000).

`threshold` uses conditional least squares to estimate the parameters of the threshold regression model. The threshold value is estimated by minimizing the SSR obtained for all tentative thresholds; see *Methods and Formulas* for details.

▷ Example 1: Threshold regression model

We are interested in the effect of inflation and the output gap on interest rates in a typical business cycle. Our dataset, `usmacro.dta`, contains quarterly data from `1954q3` to `2010q4` on the U.S. federal funds interest rate (`fedfunds`), the current inflation rate (`inflation`), and the output gap (`ogap`). These data were obtained from the Federal Reserve Economic Database, a macroeconomic database provided by the Federal Reserve Bank of Saint Louis; see [D] **import fred**.

In our model, we assume that the Federal Reserve sets the federal funds interest rate based on its most recent lag (`l.fedfunds`), the current inflation rate, and the output gap. We use the first lag of the federal funds interest rate as the threshold variable, and we assume one threshold, or two regions, so the model may be written as

$$\texttt{fedfunds}_t = \delta_{10} + \delta_{11}\texttt{l.fedfunds} + \delta_{12}\texttt{inflation} + \delta_{13}\texttt{ogap} + \epsilon_t \text{ if } -\infty < \texttt{l.fedfunds} \leq \gamma$$

$$\texttt{fedfunds}_t = \delta_{20} + \delta_{21}\texttt{l.fedfunds} + \delta_{22}\texttt{inflation} + \delta_{23}\texttt{ogap} + \epsilon_t \text{ if } \gamma < \texttt{l.fedfunds} < \infty$$

```
. use http://www.stata-press.com/data/r15/usmacro
(Federal Reserve Economic Data - St. Louis Fed)
. threshold fedfunds, regionvars(l.fedfunds inflation ogap)
> threshvar(l.fedfunds)
Searching for threshold: 1
(Running 177 regressions)
..................................................    50
..................................................   100
..................................................   150
.........................
Threshold regression
                                                Number of obs     =       222
Full sample:    1955q3 - 2010q4                 AIC               =  -63.1438
Number of thresholds =  1                       BIC               =  -35.9224
Threshold variable: L.fedfunds                  HQIC              =  -52.1535
------------------------------
Order     Threshold        SSR
------------------------------
1          9.3500        155.4266
------------------------------
```

fedfunds	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Region1						
fedfunds						
L1.	.9268958	.0356283	26.02	0.000	.8570656	.996726
inflation	.0602282	.0401287	1.50	0.133	-.0184227	.1388791
ogap	.0990296	.0234809	4.22	0.000	.0530079	.1450513
_cons	.1966223	.1447802	1.36	0.174	-.0871416	.4803863
Region2						
fedfunds						
L1.	.6974113	.0783207	8.90	0.000	.5439056	.850917
inflation	.1676449	.0540984	3.10	0.002	.061614	.2736757
ogap	.0558738	.073411	0.76	0.447	-.088009	.1997567
_cons	2.16261	.8081146	2.68	0.007	.578734	3.746485

The output consists of two tables. The first table reports the estimated threshold and the corresponding SSR. The column labeled `Order` ranks the order in which the threshold was estimated. Because there is only a single threshold in this example, the order of 1 corresponds to the threshold value that contributes most in minimizing the SSR. The order is more relevant in the case of multiple thresholds.

The estimated threshold of 9.35% splits the sample into two regions. `Region1` corresponds to the portion of the sample in which the federal funds interest rate from last quarter is less than or equal to 9.35%. `Region2` corresponds to the portion of the sample in which the federal funds interest rate from last quarter is greater than 9.35%.

Coefficient estimates appear in the second table. In `Region1`, or the low federal funds interest rate region, the coefficient of 0.93 on the lag of `fedfunds` indicates that `fedfunds` is highly persistent. The coefficient on `inflation` is not significantly different from zero, which implies that the Federal Reserve does not attach any weight to the inflation rate in the low federal funds interest rate region

and cares more about the output gap. In Region2, or the high federal funds interest rate region, the coefficient on the lag of fedfunds is only 0.70, which indicates that fedfunds is not as persistent as in Region1. In Region 2, the coefficient on ogap is not significantly different from zero, but the coefficient on inflation is, so we may infer that the Federal Reserve cares more about inflation than it does about the output gap.

◁

▷ Example 2: Selecting the threshold variable

In example 1, we use l.fedfunds as the threshold variable. The Federal Reserve may also consider the output gap to be an important factor that determines the interest rate. In this example, we fit models using the first and second lags of output gap as threshold variables. We store the estimates of each model for comparison using estimates store.

First, we store the estimates of example 1 as Model1.

```
. estimates store Model1
```

Next, we fit two models, one with l.ogap as the threshold variable and the other with l2.ogap as the threshold variable. We store the estimates as Model2 and Model3, respectively.

```
. threshold fedfunds, regionvars(l.fedfunds inflation ogap) threshvar(l.ogap)
(output omitted)
. estimates store Model2
. threshold fedfunds, regionvars(l.fedfunds inflation ogap) threshvar(l2.ogap)
(output omitted)
. estimates store Model3
```

We compare the SSR and information criteria of all fitted models. Combining all estimates, we get the following table:

```
. estimates table Model1 Model2 Model3, stats(ssr aic bic hqic)
```

Variable	Model1	Model2	Model3
Region1			
fedfunds			
L1.	.92689581	.90860624	.8533835
inflation	.0602282	.19755936	.28187753
ogap	.0990296	.29553563	.14449944
_cons	.19662232	1.4172835	.54280799
Region2			
fedfunds			
L1.	.69741126	.90512493	.90879685
inflation	.16764486	.0896271	.08361366
ogap	.05587384	.15549667	.15233276
_cons	2.1626095	.17554381	.15764634
Statistics			
ssr	155.42663	145.96457	142.0608
aic	-63.143795	-77.087586	-83.105746
bic	-35.922376	-49.866167	-55.884327
hqic	-52.153481	-66.097272	-72.115432

From the table above, we see that `Model3` provides the best fit. This is the model that uses the second lag of output gap as the threshold variable.

◁

▷ Example 3: Selecting the number of thresholds

Instead of assuming a known number of thresholds, we can use model selection to choose the number of thresholds that minimizes a certain information criterion. In example 2, using `l2.ogap` as the threshold variable provided the best fit. We fit a model with an unknown number of thresholds using `l2.ogap` as the threshold variable. We can do this by specifying the maximum number of thresholds in the `optthresh()` option. In this example, we specify 5 as the maximum number of thresholds.

```
. threshold fedfunds, regionvars(l.fedfunds inflation ogap) threshvar(l2.ogap)
> optthresh(5) nodots

Searching for threshold: 1
(Running 177 regressions)
Searching for threshold: 2
(Running 146 regressions)
Searching for threshold: 3
(Running 105 regressions)
Searching for threshold: 4
(Running 52 regressions)
Searching for threshold: 5
(Running 40 regressions)

Threshold regression

Full sample:    1955q3 - 2010q4                 Number of obs    =        222
Number of thresholds =  2                       Max thresholds   =          5
Threshold variable: L2.ogap                     BIC              =   -60.0780
```

Order	Threshold	SSR
1	-3.1787	142.0608
2	-0.5351	126.4718

fedfunds	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Region1						
fedfunds						
L1.	.8533835	.0435617	19.59	0.000	.7680042	.9387628
inflation	.2818775	.0679414	4.15	0.000	.1487148	.4150403
ogap	.1444994	.072028	2.01	0.045	.0033272	.2856717
_cons	.542808	.4297171	1.26	0.207	-.299422	1.385038
Region2						
fedfunds						
L1.	.9406721	.0338085	27.82	0.000	.8744087	1.006935
inflation	-.0191805	.0462729	-0.41	0.679	-.1098737	.0715128
ogap	.2387934	.0565521	4.22	0.000	.1279534	.3496334
_cons	.638354	.1591717	4.01	0.000	.3263832	.9503249
Region3						
fedfunds						
L1.	.8892742	.0593484	14.98	0.000	.7729535	1.005595
inflation	.1851127	.0532112	3.48	0.001	.0808206	.2894047
ogap	.1984744	.039236	5.06	0.000	.1215733	.2753754
_cons	-.3086232	.2215645	-1.39	0.164	-.7428817	.1256352

We estimate two thresholds using the default BIC (`bic`). The first estimated threshold is `l2.ogap` $= -3.18$. A negative value of `l2.ogap` implies low economic growth two quarters ago. The second estimated threshold is -0.54 and also represents a negative output gap, although with a smaller magnitude. The two thresholds split the sample into three regions.

In the first region, `Region1`, the second lag of output gap is less than or equal to -3.18, indicating a recession period. In this case, the coefficients on `inflation` and `ogap` are both significantly different from zero, which implies that the Federal Reserve considers the current inflation rate and the output gap as important predictors of federal funds interest rate.

In the second region, `Region2`, the second lag of output gap is between -3.18 and -0.54. In this case, only the coefficient on output gap is significantly different from zero, which implies that the Federal Reserve only considers the output gap as a predictor of federal funds interest rate.

In the third region, `Region3`, the second lag of output gap is greater than -0.54. In this case, the coefficients on current inflation rate and output gap are both significantly different from zero, which implies that the Federal Reserve considers the current inflation rate and the output gap as important predictors of federal funds interest rate.

◁

Stored results

`threshold` stores the following in `e()`:

Scalars

`e(N)`	number of observations
`e(k)`	number of parameters
`e(nthresholds)`	number of thresholds
`e(optthresh)`	number of maximum thresholds; if specified
`e(ssr)`	sum of squared residuals of the model
`e(rank)`	rank of `e(V)`
`e(aic)`	Akaike information criterion
`e(bic)`	Bayesian information criterion
`e(hqic)`	Hannan–Quinn information criterion
`e(tmin)`	minimum time
`e(tmax)`	maximum time

Macros

`e(cmd)`	`threshold`
`e(cmdline)`	command as typed
`e(eqnames)`	names of equations
`e(depvar)`	name of dependent variable
`e(regionvars)`	list of region-specific variables
`e(indepvars)`	list of region-invariant variables
`e(threshvar)`	name of the threshold variable
`e(criteria)`	information criteria if `optthresh(#,` *ictype*`)` is specified
`e(title)`	title in estimation output
`e(tsfmt)`	format for the current time variable
`e(tmins)`	formatted minimum time
`e(tmaxs)`	formatted maximum time
`e(vce)`	*vcetype* specified in `vce()`
`e(vcetype)`	title used to label Std. Err.
`e(properties)`	`b V`
`e(predict)`	program used to implement `predict`
`e(marginsnotok)`	predictions disallowed by `margins`
`e(asbalanced)`	factor variables `fvset` as `asbalanced`
`e(asobserved)`	factor variables `fvset` as `asobserved`

Matrices

`e(b)`	coefficient vector
`e(Cns)`	constraints matrix
`e(V)`	variance–covariance matrix of the estimators
`e(ssrmat)`	sum of squared residuals for each estimated threshold
`e(thresholds)`	matrix of estimated thresholds
`e(nobs)`	number of observations in each region

Functions

`e(sample)`	marks estimation sample

Methods and formulas

Methods and formulas are presented under the following headings:

Introduction

Consider a threshold regression with two regions defined by a threshold γ. This is written as

$$\begin{aligned} y_t &= \mathbf{x}_t\boldsymbol{\beta} + \mathbf{z}_t\boldsymbol{\delta}_1 + \epsilon_t \qquad \text{if} \quad -\infty < w_t \leq \gamma \\ y_t &= \mathbf{x}_t\boldsymbol{\beta} + \mathbf{z}_t\boldsymbol{\delta}_2 + \epsilon_t \qquad \text{if} \quad \gamma < w_t < \infty \end{aligned}$$

where y_t is the dependent variable, $\mathbf{x}_t$ is a $1 \times k$ vector of covariates possibly containing lagged values of y_t, $\boldsymbol{\beta}$ is a $k \times 1$ vector of region-invariant parameters, $\mathbf{z}_t$ is a vector of exogenous variables with region-specific coefficient vectors $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$, w_t is a threshold variable that may also be one of the variables in $\mathbf{x}_t$ or $\mathbf{z}_t$, and ϵ_t is an IID error with mean 0 and variance σ^2,

The estimated threshold ($\widehat{\gamma}$) is one of the values in the threshold variable w_t. To estimate the threshold, we minimize the least squares of the following regression with T observations and two regions,

$$y_t = \mathbf{x}_t\boldsymbol{\beta} + \mathbf{z}_t\boldsymbol{\delta}_1 I(-\infty < w_t \leq \gamma) + \mathbf{z}_t\boldsymbol{\delta}_2 I(\gamma < w_t < \infty) + \epsilon_t$$

for a sequence of T_1 values in w_t, where $T_1 < T$. The default trimming percentage is set to 10%, which implies that T_1 corresponds to the number of observations between the 10th and the 90th percentile of w_t. The estimator for the threshold is

$$\widehat{\gamma} = \arg\min_{\gamma \in \Gamma} S_{T_1}(\boldsymbol{\gamma})$$

where $\Gamma = (-\infty, \infty)$,

$$S_{T_1}(\boldsymbol{\gamma}) = \sum_{t=1}^{T} \{y_t - \mathbf{x}_t\boldsymbol{\beta} - \mathbf{z}_t\boldsymbol{\delta}_1 I(-\infty < w_t \leq \gamma) - \mathbf{z}_t\boldsymbol{\delta}_2 I(\gamma < w_t < \infty)\}^2$$

is a $T_1 \times 1$ vector of SSR, and $\boldsymbol{\gamma}$ is a $T_1 \times 1$ vector of tentative thresholds.

Model with more than two regions

In general, a threshold regression model with m thresholds has $m+1$ regions. Let $j = 1, \ldots, m+1$ index the regions. We can write the model as

$$\begin{aligned} y_t &= \mathbf{x}_t\boldsymbol{\beta} + \mathbf{z}_t\boldsymbol{\delta}_1 I_1(\gamma_1, w_t) + \cdots + \mathbf{z}_t\boldsymbol{\delta}_{m+1} I_{m+1}(\gamma_{m+1}, w_t) + \epsilon_t \\ y_t &= \mathbf{x}_t\boldsymbol{\beta} + \sum_{j=1}^{m+1} \mathbf{z}_t\boldsymbol{\delta}_j I_j(\gamma_j, w_t) + \epsilon_t \end{aligned}$$

where $\gamma_1 < \gamma_2 < \cdots < \gamma_m$ are ordered thresholds with $\gamma_0 = -\infty$ and $\gamma_{m+1} = \infty$. $I_j(\gamma_j, w_t) = I(\gamma_{j-1} < w_t \leq \gamma_j)$ is an indicator for the jth region. Conditional on all estimated thresholds $(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_m)$, the threshold regression model is linear, and the remaining parameters are estimated using least squares.

The thresholds are estimated sequentially as described below. Let $\gamma_1^*, \ldots, \gamma_m^*$ represent the m thresholds in the order of estimation. Gonzalo and Pitarakis (2002) show that the thresholds estimated sequentially are T consistent. The first threshold (γ_1^*) is estimated assuming a model with two regions as described in the previous section. Conditional on the first threshold, the second threshold is estimated as the value that yields the minimum sum of squared errors over all observations in w_t excluding the first threshold. The estimator of the second threshold γ_2^* is obtained by minimizing the least squares of a regression with three regions conditional on the first estimated threshold $\widehat{\gamma}_1^*$. The estimator is given by

$$\widehat{\gamma}_2^* = \arg\min_{\gamma_2^* \in \Gamma_2} S_{T_2}(\gamma_2^* | \widehat{\gamma}_1^*)$$

where $\Gamma_2 = (\gamma_0, \widehat{\gamma}_1^*) \cup (\widehat{\gamma}_1^*, \gamma_3)$ and $T_2 < T_1$.

In general, the lth threshold minimizes the SSR conditional on the $l-1$ estimated thresholds and is given by

$$\widehat{\gamma}_l^* = \arg\min_{\gamma_l^* \in \Gamma_l} S_{T_l}(\gamma_l^* | \widehat{\gamma}_1^*, \ldots, \widehat{\gamma}_{l-1}^*)$$

where $\Gamma_l = (\gamma_0, \gamma_{m+1})$ excluding $\widehat{\gamma}_1^*, \ldots, \widehat{\gamma}_{l-1}^*$.

When the number of thresholds is not known a priori, `threshold` selects the optimal number of thresholds based on AIC, BIC, or HQIC, which is defined based on SSR from the fitted model as

$$\begin{aligned}
\text{AIC} &= T \ln(\text{SSR}/T) + 2k \\
\text{BIC} &= T \ln(\text{SSR}/T) + k \ln(T) \\
\text{HQIC} &= T \ln(\text{SSR}/T) + 2k \ln\{\ln(T)\}
\end{aligned}$$

where k is the number of parameters in the model. See Gonzalo and Pitarakis (2002) for Monte Carlo studies of selecting the number of thresholds based on information criteria.

References

Gonzalo, J., and J.-Y. Pitarakis. 2002. Estimation and model selection based inference in single and multiple threshold models. *Journal of Econometrics* 110: 319–352.

Hansen, B. E. 1997. Approximate asymptotic p values for structural-change tests. *Journal of Business and Economic Statistics* 15: 60–67.

——. 2000. Sample splitting and threshold estimation. *Econometrica* 68: 575–603.

——. 2011. Threshold autoregression in economics. *Statistics And Its Interface* 4: 123–127.

Linden, A. 2015. Conducting interrupted time-series analysis for single- and multiple-group comparisons. *Stata Journal* 15: 480–500.

——. 2017. A comprehensive set of postestimation measures to enrich interrupted time-series analysis. *Stata Journal* 17: 73–88.

Tong, H. 1983. *Threshold Models in Non-linear Time Series Analysis*. New York: Springer.

——. 1990. *Non-linear Time Series: A Dynamical System Approach*. New York: Oxford University Press.

Also see

[TS] **threshold postestimation** — Postestimation tools for threshold

[TS] **mswitch** — Markov-switching regression models

[TS] **tsset** — Declare data to be time-series data

[R] **regress** — Linear regression

[U] 20 Estimation and postestimation commands