UNIVERSITY OF EXETER

MAX-PLANCK-INSTITUT FÜR DEMOGRAFISCHE FORSCHUNG

MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH

Efficient Programming in Stata and Mata II: Obtaining Non-Standard Distributions for a Cointegration Test via Simulation

Sebastian Kripfganz
University of Exeter Business School
Daniel C. Schneider
Max Planck Institute for Demographic Research

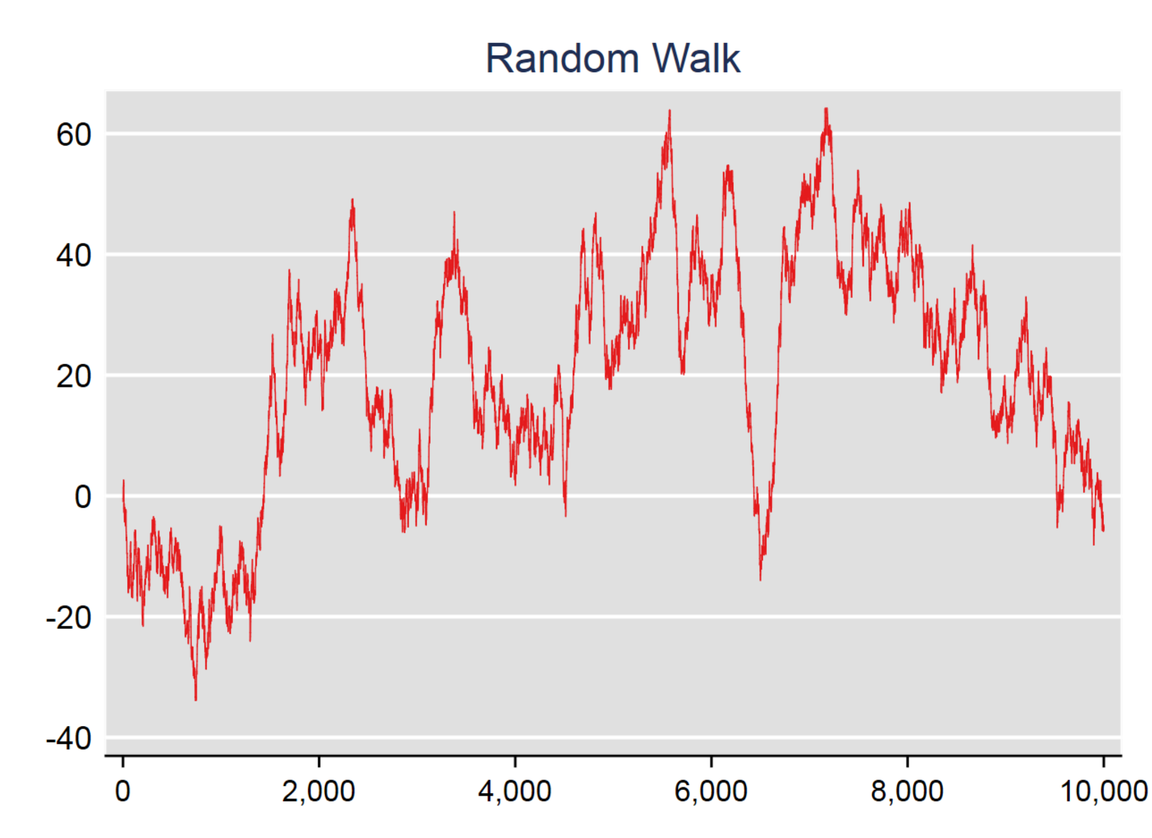
German Stata Users Group Meeting, June 22, 2018, Konstanz

Last Year's Talk

- efficient coding strategies:
  - use common sense
  - use your knowledge of your software (Stata, of course!)
  - use your knowledge of matrix algebra
- case study: the -ardl- estimation command
  - last year: optimal lag selection
  - this talk: simulation of finite sample distributions

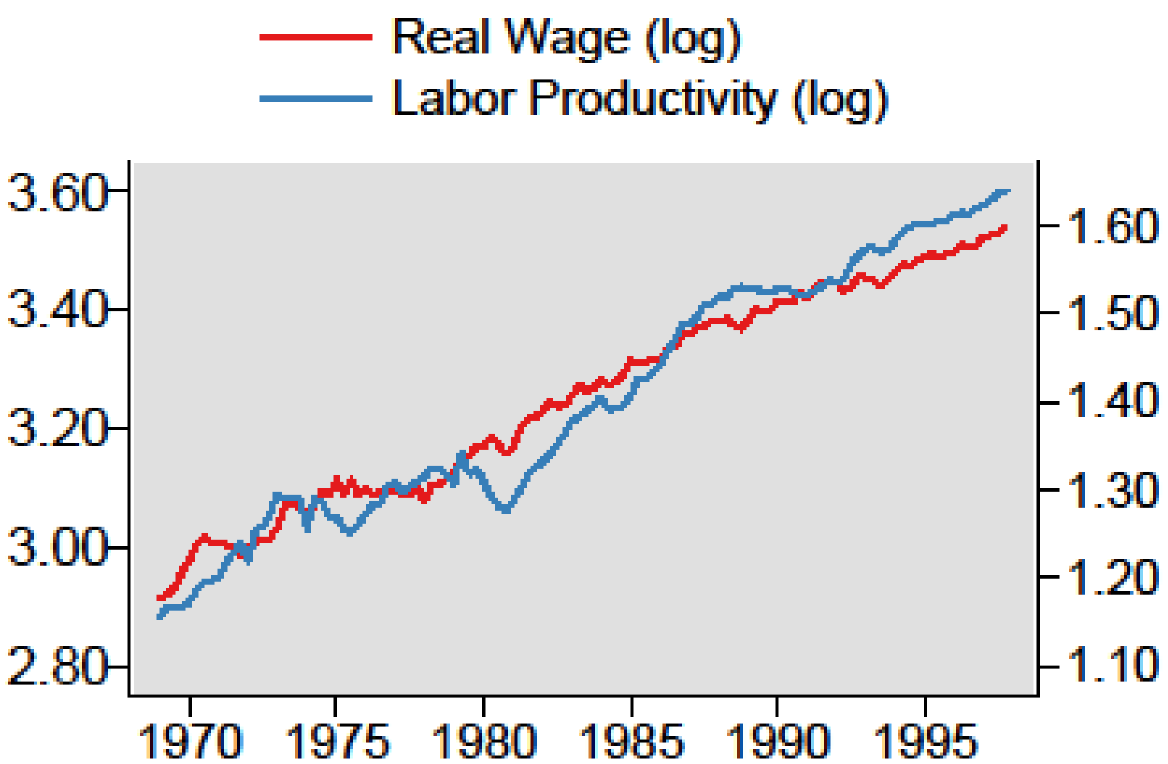
Stationarity vs. Non-Stationarity

- fundamental distinction in time series analysis (TSA)
- mostly about time series with a unit root: I(0) vs. I(1)
- non-stationary TS behave fundamentally different



Multiple Time Series Analysis

Long-run relationship: Some time series are bound together due to equilibrium forces even though the individual time series might move considerably.



Data source: Pesaran, Shin, and Smith (2001).

The ARDL Model and the Bounds Test

- ARDL$(p, q, \ldots, q)$ model:

$$y_t = c_0 + c_1 t + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=0}^{q} \boldsymbol{\beta}_i' \mathbf{x}_{t-i} + \varepsilon_t,$$

  with $\mathbf{x}_t$ a $K \times 1$ vector.

- Reparameterization in error-correction (EC) form:

$$\begin{aligned} \Delta y_t = c_0 + c_1 t - \alpha(y_{t-1} - \boldsymbol{\theta}\mathbf{x}_{t-1}) \\ + \sum_{i=1}^{p-1} \psi_{yi} \Delta y_{t-i} + \boldsymbol{\omega}' \Delta \mathbf{x}_t + \sum_{i=1}^{q-1} \boldsymbol{\psi}_{xi}' \Delta \mathbf{x}_{t-i} + \varepsilon_t, \end{aligned}$$

- Pesaran / Shin / Smith (2001) (PSS) derive the asymptotic coefficient distributions under the opposing assumptions of stationary vs. non-stationary regressors, the basis for their bounds test for a levels relationship.
- They provide critical values (CV) tables obtained via simulation.

ARDL Toy Model Estimation

```
. ardl w prod union ur , ec maxlag(6) dots trend(qtime) restricted vsquish
```

```
Optimal lag selection, % complete:
----+--20%--+--40%--+--60%--+--80%--+--100%
.............................................
BIC optimized over 2058 lag combinations

ARDL(2,0,2,0) regression

Sample: 1971q3 - 1997q4                      Number of obs     =        106
                                             R-squared         =     0.2637
                                             Adj R-squared     =     0.2029
Log likelihood =   330.70424                  Root MSE          =     0.0112
```

D.w	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ADJ						
w						
L1.	-0.240	0.063	-3.827	0.000	-0.365	-0.116
LR						
prod	0.416	0.208	1.998	0.049	0.003	0.829
union	-0.210	0.235	-0.893	0.374	-0.676	0.256
ur	0.039	0.017	2.382	0.019	0.007	0.072
qtime	0.003	0.001	2.962	0.004	0.001	0.005
SR						
w						
LD.	-0.203	0.094	-2.161	0.033	-0.389	-0.017
union						
D1.	0.058	0.597	0.097	0.923	-1.128	1.243
LD.	-1.535	0.596	-2.574	0.012	-2.719	-0.351
_cons	0.527	0.153	3.454	0.001	0.224	0.830

ARDL Toy Model Estimation

```
. estat btest

note:  estat btest has been superseded by  estat ectest
       as the prime procedure to test for a levels relationship.
       (click to run)

Pesaran/Shin/Smith (2001) ARDL Bounds Test
H0: no levels relationship                F =      3.863
                                          t =     -3.827

Critical Values (0.1-0.01),  F-statistic, Case 4
```

	[I_0] L_1	[I_1] L_1	[I_0] L_05	[I_1] L_05	[I_0] L_025	[I_1] L_025	[I_0] L_01	[I_1] L_01
k_3	2.97	3.74	3.38	4.23	3.80	4.68	4.30	5.23

```
accept if F < critical value for I(0) regressors
reject if F > critical value for I(1) regressors

Critical Values (0.1-0.01),  t-statistic, Case 4
```

	[I_0] L_1	[I_1] L_1	[I_0] L_05	[I_1] L_05	[I_0] L_025	[I_1] L_025	[I_0] L_01	[I_1] L_01
k_3	-3.13	-3.84	-3.41	-4.16	-3.65	-4.42	-3.96	-4.73

```
accept if t > critical value for I(0) regressors
reject if t < critical value for I(1) regressors

k: # of non-deterministic regressors in long-run relationship
Critical values from Pesaran/Shin/Smith (2001)
```

Simulation Project Outline

- PSS bounds test very popular, but CV tables only cover a limited number of cases

$\Rightarrow$computational / simulation project:

1. simulate distributions for all combinations of c, l, k, q, T
2. store calculated statistics / distributions
3. run response surface regressions (RSR), where the depvars are distributional quantiles
4. implement and distribute an ARDL postestimation feature that displays RSR-based CVs / p-values

Response Surface Regressions (RSR)

- idea:
  for each c, l, k: regress quantile of distr ~ g(T,q)
  We implement variations thereof.
- use predicted values for a particular T, q as CVs in applied work
- introduced by MacKinnon (1991, 1994, 1996)
- Other Stata commands, e.g.
  - `ersur` (Baum/Otero 2017)
  - `kssur`, `ksur` (Otero/Smith 2017)

The Computational Task

Similar to PSS, the DGP is

$$y_t = y_{t-1} + \epsilon_{yt}$$
$$\boldsymbol{x}_t = \boldsymbol{P}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_{xt}$$

for $t = 1, 2, \ldots, T + 50$ (including 50 burn-in periods), and where

$$(y_0, \boldsymbol{x}'_0)' = \mathbf{0}, \epsilon_t \sim N(0, I_{k+1})$$

and

$$\boldsymbol{P} = 0 \;\; (I(0) \text{ regressors})$$
$$\boldsymbol{P} = \boldsymbol{I_k} \;\; (I(1) \text{ regressors})$$

The Computational Task

project size:

Symbol	Meaning	Values	# values
c	deterministics cases	1, 2, …, 5 (F); 1, 3, 5 (t)	8
I	integration order	0, 1	2
k	# of regressors	0, 1, …, 10	11
q	# of lags	0, 1, …, 4, 6, 8, 12	8
T	sample size	20, 22, …, 400, 500, 1000	18
r	# replications		100,000
m	# meta replications		100

Results in ~160,000,000,000 stats

Implies several months of computation (“Oh my!”)

Implies ~600GB disk space (“Oh dear!”)

Reducing Data Size

Idea, omitting details: i) round to 3 decimal places, ii) store tabulation

cIkr_group	stat	stat3	tmpdif	statdif	mult
		(...)			
2310	2.345145	2345	2345	-28655	1
2310	2.761234	2761	416	-30584	2
2310	2.761411	2761	0	-31000	2
2310	2.761932	2762	1	-30999	4
2310	2.761944	2762	0	-31000	4
2310	2.761948	2762	0	-31000	4
2310	2.762331	2762	0	-31000	4
		(...)			
2310	10.85794	10858	100	-20142	1
2310	10.99043	10990	132	-20010	1
		(...)			
2311	2.118192	2118	2118	-28882	1
2311	2.239101	2239	121	-30879	1
2311	2.241233	2241	2	-30998	1
2311	2.241708	2242	1	-30999	2
2311	2.241744	2242	0	-31000	2

statdif	mult	first
	(...)	
-28655	1	1
-30584	2	0
-30999	4	0
	(...)	
-20142	1	0
-20010	1	0
	(...)	
-28882	1	1
-30879	1	0
-30998	1	0
-30999	2	0
	(...)	

Reducing Data Size

- Achieved size reduction: over 90%
- After -zipfile-, data occupy 10GB
- Solving this was crucial as now computational steps can be separated.
- But: Takes up 20% computation time
- . help data types, . help compress
- Data transformations and data types
  - Years, age in years
- Wish list item: if Mata supported all numeric types of Stata
  - Could implement more complex storage ideas in Mata and its mmat files
  - Could write (de-)compression in terms of a class

Simulation & Multiple Stata Instances

```
// --------------- beg dosim.do -----------
args inputarg
if "`inputarg'"!="" {
     confirm integer number `inputarg'
     // (...) potentially some setup statements here
     //   like startup scripts that set matsize, maxvar, etc.
}

set rng mt64s
local laglist 1 2 3 4 6 8 12
if "`inputarg'"!="" local laglist `inputarg'

foreach lag of local laglist {
     set seed 123456
     set rngstream `lag'
     mata : dosim(`lag')
}
// --------------- end dosim.do -----------
```

Simulation & Multiple Stata Instances

Windows / DOS batch file to fire up Stata instances

```
rem -------- beg multiinstance.bat ---------
for %%c in (1 2 3 4 6 8 12) do (
    copy dosim.do dosim_multiinst_ %%c.do /Y
    start "sim%%c" /D "PROJECTPATH" "STATAPATH\StataMP-64.exe" ^
          /e do dosim_multiinst_ %%c.do   %%c
)
rem -------- end multiinstance.bat ---------
```

Simulation & Multiple Stata Instances

- Multiple instances
  - help entry: [GSW] B.5 Stata batch mode
  - careful with any kind of file saving operations, e.g. logs
  - batch file to kill processes?
- RNG streams
  - new in Stata 15
  - . help set rngstream

Mata Code Optimization

- necessary to examine each expression for speed improvements
- examples of smaller improvements
  - row extraction instead of column extraction
  - inner vector product: sum of squares vs. cross() vs. multiplication
- most important code features
  - pre-calculation of cross-products, accessing through indexing
  - use pointer variables to facilitate storing numbers
  - experiment with inverters / solvers
- not pursued: C/C++
  - Stata/Mata has a MUCH better convenience-speed trade-off
  - Stata/Mata great in other respects too: version control

Mata Code Optimization

Usage of pointer variables

```
/*
Structure of returned results:

   pstatlidx                      pFkI , ptkI                     unnamed but referenced matrices
(returned matrix)

       | lag-idx                    | I                                     |  c
       | 0 1 ...                    | 0  1                                  |  1 (2) 3 (4) 5
---------------- p point to:   -----------   p point to:   -----------------------------
stat=F | p p ...                k 0 | p  p                 statdata 1    |  #  #  #  #  #
stat=t | p p ...                  1 | p  p                          2    |  #  #  #  #  #
                                 ...| ...                          ...   |  ...
                                kmax| p  p                         reps  |  #  #  #  #  #
*/
```

Mata Code Optimization

Loop structure

```
for [T] {
  for [lags] {
    // - calculate deterministics for all cases (X1)
    //   cross products thereof (XX11)
    for [reps] {
      // - random draws
      // - calculation of levels variables (X2)
      //   cross products thereof (XX22)
      // - calculation of first-difference variables (X3))
      //   cross products thereof (XX33)
      // - also calculate cross products among y, X1, X2, X3 variables (XX12, ...)
      for [cases] {
        for [k] {
          // - check degree-of-freedom requirement
          for [I-order] {
            // - select / assemble matrices from parts for (un-)restricted models (F-test)
            //   calculate (un-)restricted SSR (solver: lusolve())
} } } } } }
```

Project Results: ARDL Toy Example

```
. quietly ardl w prod union ur , ec maxlag(6) dots trend(qtime) restricted  vsquish

. estat ectest

Pesaran, Shin, and Smith (2001) bounds test

H0: no level relationship                                      F =       3.863
Case 4                                                             t =      -3.827

Finite sample ( 3 variables,  106 observations,  3 short-run coefficients)

Kripfganz and Schneider (2018) critical values and approximate p-values

   | 10%               | 5%                | 1%                | p-value
   |    I(0)     I(1)  |    I(0)     I(1)  |    I(0)     I(1)  |    I(0)     I(1)
---+-------------------+-------------------+-------------------+------------------
 F |   3.011    3.829  |   3.486    4.373  |   4.530    5.548  |   0.028    0.096
 t |  -3.116   -3.829  |  -3.419   -4.162  |  -4.016   -4.803  |   0.017    0.100

do not reject H0 if
    both F and t are closer to zero than critical values for I(0) variables
      (if p-values > desired level for I(0) variables)
reject H0 if
    both F and t are more extreme than critical values for I(1) variables
      (if p-values < desired level for I(1) variables)
```

Project Results: ARDL Toy Example

PSS values

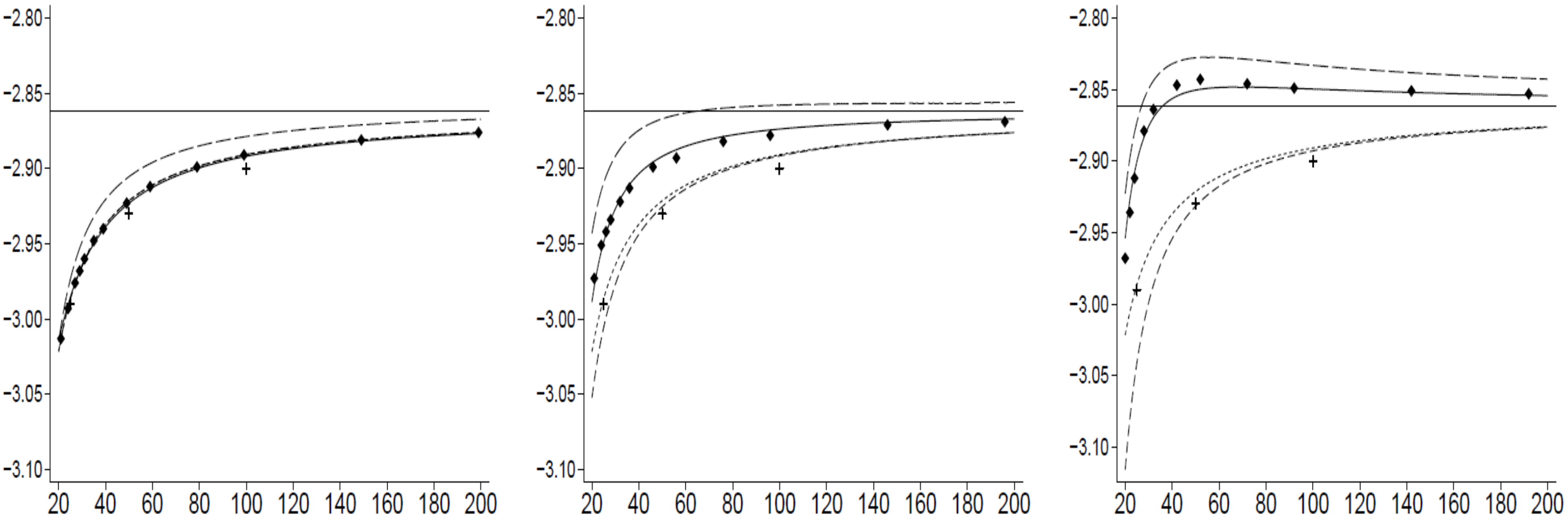
	[I_0] L_1	[I_1] L_1	[I_0] L_05	[I_1] L_05	[I_0] L_025	[I_1] L_025	[I_0] L_01	[I_1] L_01
k_3	2.97	3.74	3.38	4.23	[redacted]	[redacted]	4.30	5.23
k_3	-3.13	-3.84	-3.41	-4.16	[redacted]	[redacted]	-3.96	-4.73

Response surface regression based values

	10% I(0)	10% I(1)	5% I(0)	5% I(1)	1% I(0)	1% I(1)	p-value I(0)	p-value I(1)
F	3.011	3.829	3.486	4.373	4.530	5.548	0.028	0.096
t	-3.116	-3.829	-3.419	-4.162	-4.016	-4.803	0.017	0.100

Project Results: E.g. Dickey-Fuller

Besides Cheung and Lai (1995), the existing literature largely neglects the lag-order dependence of the finite-sample critical values (t-statistic, k=0, case (iii), α=5%)



(a) $q = 1$ (b) $q = 4$ (c) $q = 8$

Recap

- Non-stationary time series and cointegration, ardl and the PSS bounds test
- Simulation project: Improve CV tables for bounds test
  - Storing large quantity of numbers
  - Computation time
    - Multiple Stata instances
    - Code improvements within Mata

Thank you!

Questions? Comments?

schneider@demogr.mpg.de

See also: the ardl discussion thread on the Stata Forum

```
. net install ardl, from(http://www.kripfganz.de/stata/)
```

Paper available at http://www.kripfganz.de/research/index.html

References

Cheung, Y.-W. and K. S. Lai (1995a). Lag order and critical values of the augmented Dickey-Fuller test. Journal of Business & Economic Statistics 13 (3), 277-280.

Kripfganz, S. and D. C. Schneider (2018). Response Surface Regressions for Critical Value Bounds and Approximate p-values in Equilibrium Correction Models. Manuscript, University of Exeter and Max Planck Institute for Demographic Research. Available at www.kripfganz.de/research/Kripfganz_Schneider_ec.html.

MacKinnon, J. G. (1991). Critical values for cointegration tests. In R. F. Engle and C. W. J. Granger (Eds.), Long-Run Economic Relationships: Readings in Cointegration, Chapter 13, pp. 267-276. Oxford: Oxford University Press.

MacKinnon, J. G. (1994). Approximate asymptotic distribution functions for unit-root and cointegration tests. Journal of Business & Economic Statistics 12 (2), 167-176.

MacKinnon, J. G. (1996). Numerical distribution functions for unit root and cointegration tests. Journal of Applied Econometrics 11 (6), 601-618.

Otero, J. and C. F. Baum (2017). Response surface models for the Elliott, Rothenberg, and Stock unit-root test. Stata Journal 17 (4), 985-1002.

Otero, J. and J. Smith (2017). Response surface models for OLS and GLS detrending-based unit-root tests in nonlinear ESTAR models. Stata Journal 17 (3), 704-722.

Pesaran, M. H., Y. Shin, and R. J. Smith (2001). Bounds testing approaches to the analysis of level relationships. Journal of Applied Econometrics 16 (3), 289-326.