Bounds testing when the controls carry trends

Dr Merwan Roudane

07 septembre 2026

Contents

1 What this package is for

`ardldml` implements **DML-Bounds**, the procedure of Villena (2026), for testing whether two series share a long-run relationship *given* a control set that may be large and may itself be non-stationary.

The classical Autoregressive Distributed Lag (ARDL) bounds test of Pesaran, Shin and Smith (2001) works inside a conditional error-correction model

$$\Delta Y_t = a_0 + \rho Y_{t-1} + \theta D_{t-1} + \delta \Delta D_t + \sum_i \gamma_i \Delta Y_{t-i} + e_t,$$

with the joint null $H_0 : \rho = \theta = 0$ on the lagged levels. Because the integration order of D is unknown, the test *brackets* it: the lower critical value treats D as $I(0)$, the upper as $I(1)$, and the statistic is read against the interval. Outside it the verdict is conclusive; inside it, inconclusive.

1.1 The second unknown

Now condition on a large control set W_t. Projecting the lagged levels onto controls that themselves carry stochastic trends can **absorb** part of the long-run variation that identifies the error-correction relation. What then governs the null is not the integration order of the original regressors but the number of stochastic trends that survive residualisation. The bracket reappears, over a different quantity.

```r
plot_bracket(k = 10, k_tilde = 6)
```

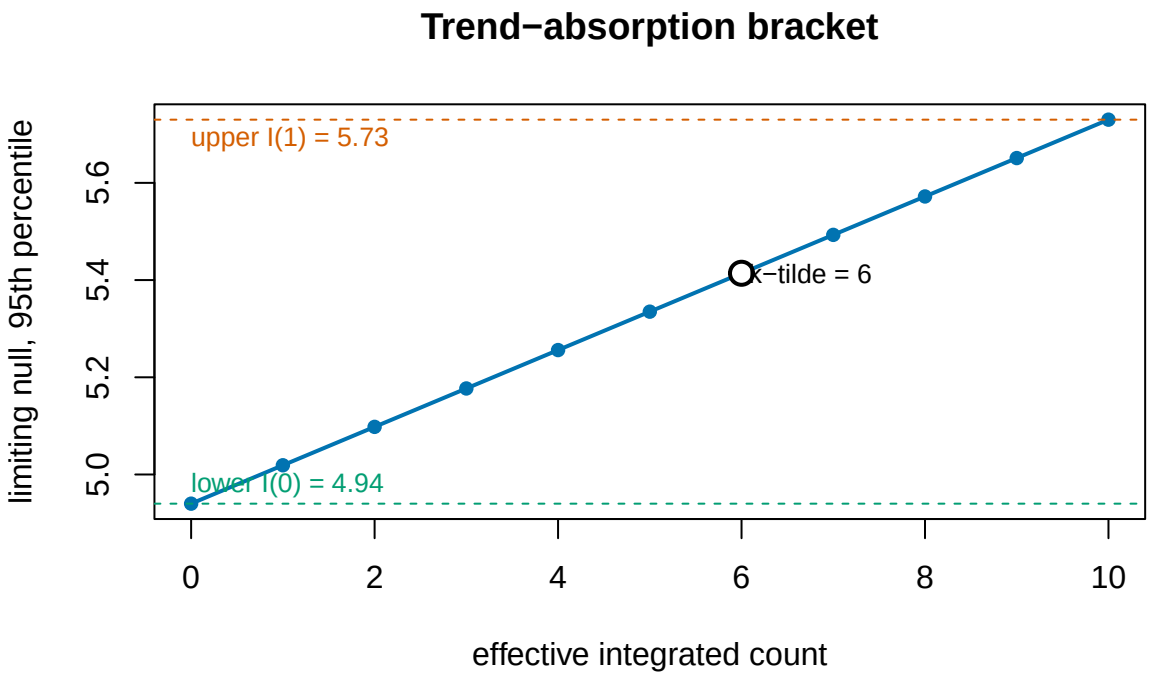

Figure 1: As the effective integrated count falls, the limiting null slides from the I(1) endpoint to the I(0) endpoint. Classical bounds testing is the right-hand end.

One consequence is worth stating plainly, because it decides how much of this you need to care about:

> **Stationary controls are harmless.** A stationary regressor cannot track the stochastic trend of an integrated one, so it cannot absorb one. Add a hundred and the effective count is unchanged. Only *integrated* controls make the bracket live.

1.2 Why no bounds table ships with this package

Tabulated critical values are not valid in this setting, and the generated-regressor remainder is not negligible at the sample sizes applied work actually uses. Every critical value here is computed instead: by simulation for the classical bracket, and by a restricted system wild bootstrap for inference on real data.

2 The data

The package bundles nine monthly United States macroeconomic series from FRED-MD (McCracken and Ng, 2016), 1973-01 to 2020-12, so every example runs offline. The application is exchange-rate pass-through: does the dollar pass through to consumer prices in the long run, conditional on the macro-financial environment?

```r
df <- passthrough_regime("1999-2007")
dim(df)
#> [1] 108  10
CONTROLS
#> [1] "m2"     "ffr"    "ip"     "unrate" "oil"    "gs10"   "baa"
DEFAULT_INTEGRATED
#> [1] "m2"   "ip"   "oil"  "gs10" "baa"  "ffr"
```

Testing within a monetary regime rather than across the whole sample avoids conflating a structural break with a long-run relationship.

3 A first fit

```
W <- as.matrix(df[, CONTROLS])
fit <- dml_bounds(df$cpi, df$neer, W,
                  lags = 4, n_blocks = 5, buffer = 6,
                  integrated = DEFAULT_INTEGRATED)
fit
#> DML-Bounds test for a conditional long-run relationship
#> F = 0.4043 on n = 103, 7 controls (6 treated as I(1)), K = 5, h = 6
#> No bootstrap critical value attached; call dml_bootstrap().
```

Three moving parts sit behind that call.

A balanced first stage. The two projections use *different* regressors, because their targets have different integration orders. ΔY_t is stationary, so the integrated controls enter its projection in first differences; regressing a stationary target on integrated levels would be unbalanced and spurious. The tested levels $Z_{t-1} = (Y_{t-1}, D_{t-1})$ are integrated and are projected on the control *levels*. All trend absorption happens in that second projection.

```
des <- build_balanced_design(df$cpi, df$neer, W, lags = 4,
                             integrated = DEFAULT_INTEGRATED)
des
#> Balanced design: n = 103
#>   X  stationary    : 12 regressors ( 1 I(0) levels, 6 I(1) differences, plus lagged differences )
#>   Z  tested levels : Y.L1, D.L1
#>   W  control levels: 7
```

h-block cross-fitting. The sample is cut into K contiguous chronological blocks, and each is evaluated by a model trained on the others *minus an h-observation buffer*. The buffer decouples first-stage error from the evaluation-fold innovations. It costs sample, and the package makes that cost visible before you commit to a configuration:

```
round(sample_use_table(nrow(df)), 3)
#>       h=0   h=2   h=5  h=10
#> K=4 0.750 0.722 0.681 0.611
#> K=5 0.800 0.770 0.726 0.652
#> K=6 0.833 0.802 0.756 0.679
#> K=8 0.875 0.843 0.794 0.713
```

A restricted system wild bootstrap. One Rademacher weight is applied to the *stacked* pair of conditional and marginal residuals, and both series are regenerated jointly. This preserves the contemporaneous correlation between them, which is the endogeneity channel the whole framework exists for. A scheme that holds D fixed and reweights only the equation error simulates a world with zero correlation whatever the data say.

```
fit <- dml_bootstrap(fit, B = 49, seed = 20260625)
summary(fit)
#> DML-Bounds test for a conditional long-run relationship
#> H0: no level relationship after residualisation      F = 0.4043
#> n = 103, controls = 7 (6 treated as I(1)), K = 5, h = 6
#> selected: 6 stationary, 1 level controls
#>
#> restricted system wild bootstrap: B = 49, system scheme
#>   critical value (95%) = 2.2157
#>   bootstrap p-value       = 0.6122
#>   decision at 5%          -> fail to reject
#>
```

```
#> speed of adjustment  alpha = 0.0022
#> long-run coefficient theta = -0.9715  (se 0.8975)
```

```
plot(fit)
```

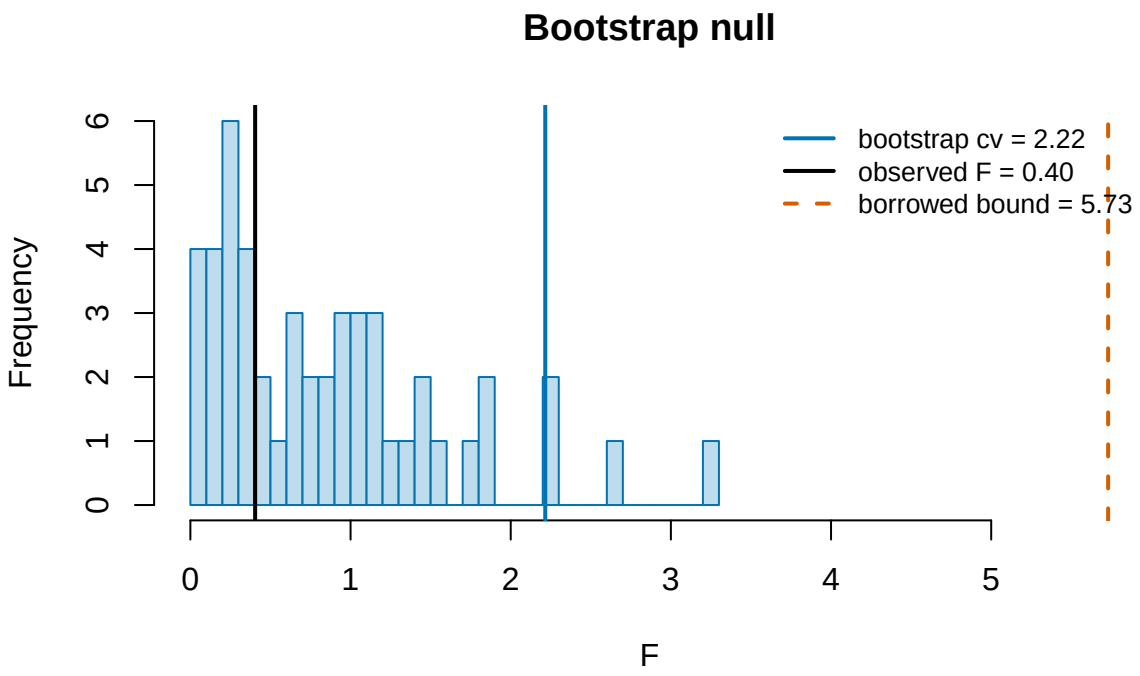

Figure 2: The bootstrap null against the borrowed classical bound. The gap between the two is the argument for not using a table.

The number of bootstrap draws is kept small here so the vignette builds quickly. Use `B = 999` for anything you intend to report.

4 The diagnostic you must run

The estimand is conditional. If a control is itself part of the equilibrium system, residualising removes the relation rather than the confounding, and a non-rejection then means nothing.

`trend_absorption()` runs four fits – full and reduced control sets crossed with the adaptive and unpenalised level projection – and reads the two gaps together with the stability of the long-run coefficient.

```
ta <- trend_absorption(df$cpi, df$neer, W, drop = REDUCED_DROP,
                       B = 19, seed = 1,
                       lags = 4, n_blocks = 5, buffer = 6,
                       integrated = DEFAULT_INTEGRATED)
ta
#> Trend-absorption diagnostic
#> dropped from reduced set: m2, oil
#>
#>  controls projection n_controls      F boot_cv95 boot_p        verdict    alpha
#>      full   adaptive          7 0.4043     3.697 0.7895 fail to reject 0.002187
#>      full        ols          7 1.1066     2.551 0.4211 fail to reject 0.018339
#>   reduced   adaptive          5 1.3467    13.421 0.5789 fail to reject 0.005664
#>   reduced        ols          5 2.3336     3.436 0.1053 fail to reject 0.002623
```

```
#>     theta theta_se
#>  -0.97148   0.8975
#>  -0.01219   0.1652
#>  -0.90428   0.3806
#>  -2.79099   7.4333
#>
#> Delta_m = p_ols  - p_ad  = -0.3684
#> Delta_W = p_full - p_red = +0.2105
#> theta spread across fits = 2.7788
#>
#> Concordant verdicts across control sets (full p = 0.789, reduced p = 0.579). The conclusion does not
```

A verdict that flips when trend-sharing controls are dropped is the signature of over-absorption, and the reduced-set verdict is then the more credible one. Concordant verdicts indicate the conclusion is not an artefact. This is a hypothesis-generating device, not a formal test: it has no size and no power, and it tells you where to look.

5 Penalty sensitivity

A verdict can turn on the penalty and on the projection. Reporting one cell is specification search; reporting the grid is the method. The column to watch is `n_selected_Z`, the number of control *levels* retained – the empirical counterpart of the effective integrated count.

```
sw <- penalty_sensitivity(df$cpi, df$neer, W,
                          lags_grid = 4, n_blocks = 5, buffer = 6,
                          integrated = DEFAULT_INTEGRATED)
sw
#>   lags projection penalty n_selected_Z      F boot_p     alpha    theta
#> 1    4   adaptive     low            1 0.5454     NA 0.0023801 -1.02062
#> 2    4   adaptive  medium            1 0.2794     NA 0.0020795 -0.91495
#> 3    4   adaptive    high            1 0.6869     NA 0.0006602 -2.84929
#> 4    4      plain     low            5 0.8890     NA 0.0145112 -0.04611
#> 5    4      plain  medium            5 1.2933     NA 0.0239062  0.02327
#> 6    4      plain    high            5 0.3238     NA 0.0092150 -0.05648
#> 7    4        ols       -            7 1.1250     NA 0.0542727  0.05914
#>   theta_se
#> 1  0.90982
#> 2  0.88789
#> 3 15.78341
#> 4  0.24634
#> 5  0.11246
#> 6  0.44456
#> 7  0.03406
attr(sw, "theta_sign_flips")
#> [1] TRUE
```

A long-run coefficient that changes sign across this grid is a warning that the conditioning set, not the data, is driving the answer.

6 Where the critical values come from

The classical bracket is regenerated from the data-generating process Pesaran, Shin and Smith printed in the notes to their Table CI, rather than transcribed. That makes it checkable against print, and extends it to any sample size and past $k = 10$, where the published tables stop.

```
sim <- simulate_pss_bounds(k = 1, case = 3, TT = 1000, nsim = 300, seed = 11)
sim
#>   level    I0    I1
#> 1  0.10 4.146 4.679
#> 2  0.05 5.039 5.564
#> 3  0.01 6.824 6.721
pss_reference(k = 1, case = 3)
#>   level   I0   I1
#> 1  0.10 4.04 4.78
#> 2  0.05 4.94 5.73
#> 3  0.01 6.84 7.84
```

With only 300 replications the simulated values carry visible Monte Carlo error; the published table used 40,000. The point of the comparison is that the generator can be checked, not that 300 draws suffice.

The classical test itself is available for benchmarking, and reads its statistic against a bracket simulated at *its own* sample size rather than one calibrated at $T = 1000$:

```
cb <- classical_bounds_test(df$cpi, cbind(neer = df$neer), lags = 4,
                            nsim = 300, seed = 11)
cb
#> Classical bounds test (Pesaran, Shin and Smith 2001)
#> H0: no level relationship           F = 1.629
#> Case 3 (unrestricted intercept, no trend)    t = -1.128
#> k = 1 forcing regressor(s), 104 observations, 2 restrictions
#>
#> simulated bounds (T = 104):
#>  level    I0    I1
#>   0.10 4.039 4.931
#>   0.05 4.617 6.080
#>   0.01 6.677 7.458
#>
#> step 3  Wald(theta = 0) = 3.883  (chi2 p = 0.0488)
#> speed of adjustment alpha = 0.0076
```

7 Validation

The statistic is deterministic given the specification, so it can be checked against an independent implementation of the same procedure. On the four monetary regimes with the full control set, this package and a separate Python implementation agree to four decimal places:

regime	this package	reference
1973–1985	9.111	9.111
1986–1998	18.712	18.712
1999–2007	0.404	0.404
2008–2020	3.703	3.703

The test suite asserts these values, along with the classical statistics and the fold structure, so a change that moves any of them fails loudly.

8 Limitations

Stated plainly, because they affect how a result should be read.

- **Power is modest at small T.** Under integrated nuisance, bootstrap power reaches useful levels only from roughly $T = 250$. Much applied ARDL work runs at $n = 30$–80, where this is not the right tool; use the classical test with simulated finite-sample bounds instead.
- **Over-absorption is not testable.** The requirement that controls span confounding trends but not the cointegrating relation cannot be verified. The diagnostic is a signal, not a guarantee.
- **The penalty can change the verdict.** Report the grid, not one cell.
- **The long-run coefficient is less reliable than the verdict.** Across regimes it is unstable and often imprecisely estimated, while the test decisions are stable.
- **The integrated block must stay small.** The validity theory keeps it fixed-dimensional; selecting among many integrated regressors is an open problem, and the package warns when you cross a threshold.

9 References

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68. doi:10.1111/ectj.12097

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4), 574–589. doi:10.1080/07350015.2015.1086655

Narayan, P. K. (2004). Reformulating critical values for the bounds F-statistics approach to cointegration. Monash University Discussion Paper 02/04.

Pesaran, M. H., Shin, Y. and Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics*, 16(3), 289–326. doi:10.1002/jae.616

Villena, M. J. (2026). Testing cointegration with many persistent controls. SSRN working paper. doi: 10.2139/ssrn.6472826

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429. doi:10.1198/016214506000000735