Chapter 9

Unit Root Testing

Despite over three decades of applied work aimed at determining whether unit roots exist in important macroeconomic variables, the question remains open in many cases. It's extremely difficult with a single time series of limited length to tell the difference between the permanent response to a shock implied by a unit root, and a response with a half-life of 20 quarters (dominant root roughly .97). We can't simply get more data by extending the front end of the data set, as the results from the literature on unit root testing and structural breaks show that breaks can cause false acceptances of unit roots. As in other cases, panel data can offer an alternative way to bring more data to bear on a questions. If we can't get a longer time span from one country, what about doing joint inference on multiple countries?

The general structure used by most (though not all) panel unit root testing procedures is:

$$\Delta y_{it} = \rho_i y_{i,t-1} + \sum_{l=1}^{p_i} \phi_{i,l} \Delta y_{i,t-l} + \alpha_i d_{it} + \varepsilon_{it} \tag{9.1}$$

where the d_{it} are the deterministic components. $\rho_i = 0$ means the y process has a unit root for individual i, while $\rho_i < 0$ means that the process is stationary around the deterministic part.

There have been quite a few possible procedures which have been proposed for testing unit roots in panel data. As part of the decision as to which to employ, we have to deal with the following questions:

1. What is the null hypothesis? In most cases, that will be unit root for all individuals.
2. What is the alternative?
3. What's heterogeneous and what (if anything) is homogeneous?
4. How do we deal with the small sample effects?

Item 2 is uninteresting in a single time series—if the null is unit root, the alternative is a stationary dominant root. In a panel setting, however, the alternative can be a single common dominant stationary root, or heterogeneous stationary roots, or even the rather vague "not all unit roots" (that is, some could have unit roots, but not all do).

Regarding item 3, the testing procedures almost uniformly allow the short-run dynamics (the lag polynomial in $\Delta y_{i,t}$) to differ among individuals, not just in coefficients, but also in the number of lags p_i. Because differing values of p_i mean different samples, the testing procedures need to allow for unbalanced samples. The coefficients on deterministic variables and the variance of ε_{it} will also generally be allowed to vary, which means that almost everything other (perhaps) than ρ_i will be heterogeneous. It's important to note that there are many perfectly reasonable ways to choose the lag length p_i which won't necessarily give the same result, particularly when applied to multiple short time series. As a result, there won't be a unique "correct" value for any test which relies upon lag pruning. This is also true if a test depends upon a long-run variance, as the value will depend upon the lag window chosen.

The small sample effect is embedded in the Dickey-Fuller and other such test statistics on single time series. A different calculation will be needed for each form of the test using panel data.

9.1 The Example

The data file from Example **9.1** is `pennxrate.dta`, which is a file with real exchange rate data on a balanced panel consisting of 151 countries observed over 34 years, from 1970 through 2003. The data are derived from the Penn World tables.[1] The U.S. is treated as the base country for exchange rates, and isn't included in the data set. The main series of interest is `LNRXRATE`, which is the log real exchange rate. The data file includes dummies for the G7 countries (there will be six of them, since the U.S. is the base) and OECD countries. If PPP holds, then there should *not* be a unit root in this series. The data file is read with

```
open data pennxrate.dta
calendar(panelobs=34,a) 1970
data(format=dta) 1//1970:01 151//2003:01 year xrate ppp id $
  capt realxrate lnrxrate oecd g7
```

9.2 Levin-Lin-Chu test

Levin, Lin, and Chu (2002) propose a test which has an alternative hypothesis that the ρ_i are identical and negative.[2] Because ρ_i is fixed across i, this is one of the most complicated of the tests because the data from the different individuals need to be combined into a single final regression. To isolate only the ρ_i in (9.1), the residuals from regressions of Δy_{it} and $y_{i,t-1}$ from all the "nuisance" variables (lags and deterministics) are obtained using individual by

[1]This is an example taken from the documentation for the Stata command `xtunitroot`.

[2]This circulated as a working paper with just Levin and Lin, so is more commonly known as the Levin-Lin test.

individual regressions. This is an application of the Frisch-Waugh Theorem to a linear regression stacked across individuals.

Each individual's data is scaled down by a feasible estimate of the standard deviation of the variance of ε_{it}. This produces two series $\tilde{e}_{i,t}$ (from Δy_{it}) and $\tilde{v}_{i,t-1}$ (from $y_{i,t-1}$). The basic test statistic is the t statistic on the linear regression of $\tilde{e}_{i,t}$ on $\tilde{v}_{i,t-1}$. For a single individual, this would be identically the Dickey-Fuller t-test statistic for the given set of augmenting lags and deterministic components—that's what the Frisch-Waugh Theorem gives us. However, we can't use the tabulated D-F distribution since both the numerator and denominator are aggregated across individuals. Instead, the authors provide centering and normalizing constants such that, as $N \rightarrow \infty$ [3] an adjusted t converges to a standard Normal. In addition to centering and normalizing constants (which depend upon (average) T and the choice of deterministics), there's also a need to correct for the different long-run variances of the $\tilde{v}_{i,t}$ processes across i. For a single time series, the long-run variance cancels out of the asymptotic distribution of the t (under the null of a unit root), but won't when both numerator and denominator are first aggregated across individuals with different short-run dynamics.

As a complicated multi-step procedure, there are a number of places where two implementations can differ. In addition to possibly different choices for lag length, both the short- and long-run variance calculations can be computed either under the null that ρ is zero, or under the alternative where ρ has been estimated. And the long-run variance depends upon choice of lag window and lag length.

The procedure **@LEVINLIN** can be used to do this test. This has quite a few options for controlling the calculation. First, the deterministic components are selected using a standard (for unit root tests) DET option:

```
DET=NONE/[CONSTANT]/TREND
```

where the default is constant (that is, individual fixed effects since the coefficients on the deterministic variables are heterogeneous). DET=TREND adds individual-specific constant and trend.

The lag length selection for the lagged differences is governed by:

LAGS=*(maximum) number of additional lags of the differenced series*
CRIT=FIXED/[GTOS]/AIC/BIC/HQ
SLSTAY=*significance level for keeping the marginal lag in* CRIT=GTOS

These again are fairly standard for RATS procedures for ADF type analysis. The default is to do general-to-specific, dropping lags as long as the significance

[3] With $N/T \rightarrow 0$.

level of the t-statistic on the final one is greater than the SLSTAY value (.10 as default).

Finally, the lag window for estimating the long-run variance is chosen using the options:

LWINDOW=[NEWEY]/BARTLETT/FLAT/PARZEN/QUADRATIC
BANDWIDTH=# *of lags (or bandwidth for* LWINDOW=QUADRATIC)

Since the LAGS option is already being used, the lags or bandwidth is chosen using the BANDWIDTH option.

In addition to these options, the **@LEVINLIN** procedure offers an alternative to compute the long-run variance. The standard calculation takes the residuals from a regression of Δy_{it} on the deterministics and applies the chosen lag window to those to get the long-run variance for an individual.[4] Under the null, however, the regression of the Δy_{it} on both the deterministics and the augmenting lags offers an estimate of the long-run variance by recoloring: the ratio of long- to short-run variances is

$$\frac{1}{\left(1-\sum_{l=1}^{p_i}\phi_{i,l}\right)^2}$$

You can choose this with the option RECOLOR, which is off by default. This has the advantage that the two variances are computed over the same range, while the standard calculation uses a longer sample range for the long-run variance (basically the full individual sample) than the short-run. It also eliminates one additional choice that could affect the test statistic.

On the example program, we use three different choices for computing the long-run variance: Bartlett (or Newey-West) windows of width 10 and 5, and the recolored AR. We use the SMPL option to restrict the sample to just the G7 countries.[5]

```
@levinlin(smpl=g7,lags=10,crit=aic,band=10) lnrxrate
@levinlin(smpl=g7,lags=10,crit=aic,band=5) lnrxrate
@levinlin(smpl=g7,lags=10,crit=aic,recolor) lnrxrate
```

The output from the final one of these is

```
Levin-Lin Unit Root Test: Series LNRXRATE
Test has large N, N/T-->0
Null is Unit Root. Alternative is Common Stationary Root.
Individual Specific Components: Constant
With average lags 1.00 chosen from 7
Long-run variances by recolored AR

N                    6
T                   34
t-unadjusted -6.73614
t-adjusted   -1.95162
Signif        0.02549
```

[4] In Baltagi (2008), Step 2 on page 276 omits the step of extracting the deterministics.
[5] The usable sample will be further restricted by lags, but that's all done internally.

The t-adjusted in the actual Levin-Lin test statistic and the significance level is the one-tailed (negative) comparison of that with a standard Normal. The bandwidth of 10 strikes me as too wide for just 34 data points. It gives a much more significant result of -3.44300. The narrower bandwidth gives -2.23023 which is more in agreement with the recolored value. While you would reject a unit root (thus accepting PPP) with any of these, we have a difference between a lukewarm rejection at a significance level of .02 vs. a rather emphatic rejection of with a p-value of 0.0003 for the wide bandwidth from different variations on the same test.

9.3 Harris-Tzavalis Test

A similar but simpler test is described by Harris and Tzavalis (1999). This also has a null of unit root versus an alternative with a single stationary value. It's designed to be applied to data sets which are relatively short in T. In order to provide relatively exact corrections for small values,[6] they very tightly restrict the model to exclude the augmenting lags. Thus if the original panel is balanced (which they require), it will remain so. They also assume a homogeneous variance which the Levin-Lin test doesn't. The test, as implemented, uses y_{it} rather than Δy_{it} as the dependent variable, which means that the test is for $\rho = 1$ rather than $\rho = 0$. It has large N, fixed T asymptotics, again, with the centered and rescaled test statistic being $N(0,1)$.

This is implemented in RATS using the procedure **@HTUNIT**, which is a relatively simple procedure since there are so few options. This applies the test to the G7 and OECD subsamples. Note that $N = 6$ for the first of these is probably too small.

```
@htunit(smpl=g7) lnrxrate
@htunit(smpl=oecd) lnrxrate
```

The output from the G7 test is:

```
Harris-Tzavalis Test: Series LNRXRATE
Test has fixed T, large N asymptotics
Null is rho(i)=1. Alternative is rho(i)==rho<>1
Individual Specific Components: Constant

N            6
T           34
Rho    0.80615
Z     -2.95168
Signif 0.00158
```

9.4 Im-Pesaran-Shin Test

Im, Pesaran, and Shin (2003) start with the same basic model (9.1), but, unlike Levin-Lin-Chu and Harris-Tzavalis, they allow the more general alternative that the ρ_i can vary and, in fact, that some individuals can have a unit root.

[6] Levin-Lin-Chu provide a rather coarse table with (average) T values of 25, 30, 35, 40 etc.

Of course, the power of the test diminishes quite severely if a substantial fraction have a unit root. With everything heterogeneous, the simplest approach is to compute separate ADF test statistics on each individual and combine those (by simple averaging of the t-statistics). The final test statistic is a normalized and rescaled version of this called $Z_{\tilde{t}-bar}$ which has an asymptotic $N(0,1)$ distribution. This has large T-large N asymptotics.

In RATS, this is performed with the **@IPSHIN** procedure. Because it does an ADF test for each individual, it has the standard options for controlling that:

LAGS=*(maximum) number of additional lags of the differenced series*
CRIT=FIXED/[GTOS]/AIC/BIC/HQ
SLSTAY=*significance level for keeping the marginal lag in* CRIT=GTOS

Note that the limit on the number of lags is 8 as that is as high as the adjustment tables go.

The authors allow only for constant or constant and trend. However, to maintain the syntax with the other tests, the DET option still reads:

DET=NONE/[CONSTANT]/TREND

but DET=NONE generates an error.

For the G7 and full OECD countries, the instructions are:

```
@ipshin(smpl=g7,lags=8,crit=aic) lnrxrate
@ipshin(smpl=oecd,lags=8,crit=aic) lnrxrate
```

The output for the G7 is

```
Im-Pesaran-Shin Unit Root Test: Series LNRXRATE
Test has large N on large T asymptotics
Null is Unit Root. Alternative is rho(i)<>1 for some i
Individual Specific Components: Constant

N                      6
T                     31 to          33
Avg P           1.000000 chosen from 8 by AIC
               Statistic Signif Level
Z tbar         -2.902291     0.001852
Z ttildebar    -2.328060     0.009954
```

9.5 Breitung Test

Breitung (2000) proposes an alternative set of procedures to Levin-Lin-Chu that use unbiased estimators rather than bias-corrected ones. First consider the case with no drift:

$$y_{it} = \alpha_i + x_{it}, \phi_i(L)x_{it} = \varepsilon_{it}$$

For simplicity, we'll assume that there are no nuisance short-run dynamics, so

$$\phi_i(L) = (1 - L) - \rho L$$

The null is $\rho = 0$ vs the alternative $\rho < 0$. (This is the same as Levin-Lin-Chu). Under the null,

$$\Delta y_{it} = \varepsilon_{it}$$

and

$$y_{it} = y_{i0} + \sum_{s=1}^{t} \varepsilon_{is}$$

If, instead of extracting the sample mean, we subtract y_{i0} from y_{it}, then

$$\widetilde{\Delta y}_{it} \equiv \Delta y_{it} = \varepsilon_{it}$$

and

$$\tilde{y}_{i,t-1} = y_{i,t-1} - y_{i0} = \sum_{s=1}^{t-1} \varepsilon_{is} \tag{9.2}$$

are uncorrelated by construction (under the null). Now assume that we have individual-specific trends:

$$y_{it} = \alpha_i + \beta_i t + x_{it}, \phi_i(L) x_{it} = \varepsilon_{it}$$

Now, again with no short-run dynamics, we have

$$\Delta y_{it} = \beta_i + \varepsilon_{it}$$

and

$$y_{it} = y_{i0} + \beta_i t + \sum_{s=1}^{t} \varepsilon_{is}$$

We can detrend y_{it} using

$$\tilde{y}_{it} = y_{it} - \left(y_{i0} + \frac{t}{T} \left(y_{iT} - y_{i0} \right) \right) = \sum_{s=1}^{t} \varepsilon_{is} - \frac{t}{T} \sum_{s=1}^{T} \varepsilon_{is}$$

However, unlike (9.2), this is a function of all ε_{is} rather than just those dated t and earlier. To produce an unbiased estimator, Δy_{it} is de-meaned using a forward operation like

$$\widetilde{\Delta y}_{it} = \Delta y_{it} - \frac{1}{T-t} \sum_{s=t+1}^{T} \Delta y_{it} = \varepsilon_{it} - \frac{1}{T-t} \sum_{s=t+1}^{T} \varepsilon_{is}$$

While both $\widetilde{\Delta y}_{it}$ and $\tilde{y}_{i,t-1}$ are functions of all ε_{is}, they are uncorrelated since the first term in $\tilde{y}_{i,t-1}$ uses only subscripts through $t-1$ and thus has zero correlation with $\widetilde{\Delta y}_{it}$ constructed with subscripts from t on, while the second is uncorrelated with $\widetilde{\Delta y}_{it}$ because the weights are equal in the sum in $\tilde{y}_{i,t-1}$ and the weights sum to zero in $\widetilde{\Delta y}_{it}$. [7]

[7] $\widetilde{\Delta y}_{it}$ is actually scaled by $\sqrt{\frac{T-t}{T-t+1}}$ to standardize the variances. That doesn't change the fact that the weights sum to zero.

In both cases, in regressing $\widetilde{\Delta y}_{it}$ on $\tilde{y}_{i,t-1}$, we have a non-stationary regressor under the null. However, in the regression over the panel, it's a sum of N independent functions of non-stationary regressors. Breitung uses N before T asymptotics to eliminate the problem with that.[8]

Short-run dynamics are handled by taking deviations from regressions onto lagged differences for the difference and the lagged dependent variable (as in Levin-Lin-Chu). However, this is done *before* the detrending/de-meaning operations.[9] $\widetilde{\Delta y}_{it}$ and $\tilde{y}_{i,t-1}$ are divided by an estimate of σ_i to correct for heterogeneous variances. The test is the t statistic from regressing the full sample (standardized) $\widetilde{\Delta y}_{it}$ on $\tilde{y}_{i,t-1}$, which has an asymptotic $N(0,1)$ distribution.

In RATS, this is performed with the procedure **@BREITUNG**. This has the standard controls for choosing individual-specific lag lengths, though the asymptotics actually rely on fixed large T, so using different numbers of lags would invalidate that. Whether that's a problem in practice isn't clear.

> LAGS=*(maximum) number of additional lags of the differenced series*
> CRIT=FIXED/[GTOS]/AIC/BIC/HQ
> SLSTAY=*significance level for keeping the marginal lag in* CRIT=GTOS

Again, the DET option has the form:

> DET=NONE/[CONSTANT]/TREND

The cases where DET=TREND is appropriate are the ones where Breitung's procedure should perform better than the Levin-Lin-Chu and Im-Pesaran-Shin tests.

Finally, there is the option

> ROBUST/[NOROBUST]

With ROBUST, the final regression uses an Eicker-White heteroscedasticity consistent covariance matrix in constructing the test statistic. With NOROBUST, it uses a standard t calculation (though without scaling by σ^2 since the data are already scaled).

For the OECD sample, the Breitung test is done with:

```
@breitung(smpl=oecd,det=constant) lnrxrate
```

with output

[8] For large N, fixed T, a central limit theorem reduces the non-stationary part to calculation of its sample mean. I'm not 100% convinced of one step in Breitung's proof, but if the final regression is done using Eicker-White standard errors, it should be fine.

[9] Levin-Lin-Chu does them at the same time.

```
Breitung Unit Root Test: Series LNRXRATE
Test has large N, large T (sequential) asymptotics
Null is Unit Root. Alternative is Common Stationary Root.
Individual Specific Components: Constant
With average lags 1.59 chosen from 9

N                  27
T                  34
Test Statistic -3.22993
Signif          0.00062
```

9.6 Hadri Test

Unlike the previous tests, the Hadri (2000) proposes a test where the null is stationarity. This is a generalization of the KPSS fluctuations test (Kwiatkowski, Phillips, Schmidt, and Shin (1992)) for a single time series. If the residuals from the deterministic part of the series form a stationary process, the partial sums of the residuals (properly scaled) form a *Brownian Bridge*. If the residuals are non-stationary, those same partial sums should have more extreme values than would be compatible with a Brownian Bridge. The typical test statistic for univariate time series is

$$\frac{1}{T^2\psi^2}\sum_{t=1}^{T} S_t^2$$

where ψ^2 is the long-run variance of the residual process. The panel test statistic aggregates across i, centers and normalizes to create an asymptotically $N(0,1)$. It rejects in the right-tail (fluctuations too large).

In RATS, this is done using the **@HADRI** procedure. This has the usual DET option to choose the deterministic components. Again, NONE is included, but isn't a valid option.

DET=NONE/[CONSTANT]/TREND

There are three forms of test statistic, which govern how the scaling coefficients ψ_i^2 are determined. If the residuals are assumed to be serially uncorrelated, they can be either be the same or different across individuals. Or, you can allow for serial correlation. You choose among the three using the option:

VARIANCE=[HOMOGENEOUS]/HETEROGENEOUS/ROBUST

For the calculations robust to serial correlation, the ψ_i^2 are computed as separate long-run variance estimates across individuals. The method of computing the long-run variance is chosen using the options:

LWINDOW=[NEWEY]/BARTLETT/FLAT/PARZEN/QUADRATIC
LAGS=# *of lags (or bandwidth for* LWINDOW=QUADRATIC)

just as they would be for HAC standard error calculations on a **LINREG** instruction. For the OECD data, this does one test using homogeneous variances, and one with long-run variance calculations.

```
@hadri(smpl=oecd,det=constant) lnrxrate
@hadri(smpl=oecd,det=constant,variance=robust,$
   lwindow=bartlett,lags=5) lnrxrate
```

The output from the second is:

```
Hadri LM Unit Root Test: Series LNRXRATE
Test has large N, large T asymptotics
Null is Stationary. Alternative is Some Unit Roots
Individual Specific Components: Constant
Robust to Serial Correlation, Bartlett(Newey-West)(5)

N           27
T           34
Z     6.932146
Signif 0.000000
```

Note that this gives the opposite conclusion from the other tests—since the null is stationarity, rejection means that we find fairly strong evidence of non-stationarity. All of the tests have been based upon independence across individuals. If there is, for instance, a common time component, that wouldn't be true. We can repeat tests with deviations from means at each time period with:

```
panel(entry=1.0,time=-1.0,smpl=oecd) lnrxrate / cxrate
@levinlin(smpl=oecd,lags=10,crit=aic,band=5) cxrate
@htunit(smpl=oecd) cxrate
@ipshin(smpl=oecd,lags=8,crit=aic) cxrate
@hadri(smpl=oecd,det=constant,variance=robust,$
   lwindow=bartlett,lags=5) cxrate
```

However, this doesn't change the problem with conflicting results.

Example 9.1 Panel Unit Root Tests

This implements the different unit root testing procedures described in this chapter.

```
open data pennxrate.dta
calendar(panelobs=34,a) 1970
data(format=dta) 1//1970:01 151//2003:01 year xrate ppp id $
  capt realxrate lnrxrate oecd g7
*
* Levin-Lin-Chu test (G7 only)
*
@levinlin(smpl=g7,lags=10,crit=aic,band=10) lnrxrate
@levinlin(smpl=g7,lags=10,crit=aic,band=5) lnrxrate
@levinlin(smpl=g7,lags=10,crit=aic,recolor) lnrxrate
*
* Harris-Tzavalis test
*
@htunit(smpl=g7) lnrxrate
@htunit(smpl=oecd) lnrxrate
*
* Im-Pesaran-Shin test
*
@ipshin(smpl=g7,lags=8,crit=aic) lnrxrate
@ipshin(smpl=oecd,lags=8,crit=aic) lnrxrate
*
* Breitung test
*
@breitung(smpl=oecd,det=constant) lnrxrate
*
* Hadri test
*
@hadri(smpl=oecd,det=constant) lnrxrate
@hadri(smpl=oecd,det=constant,variance=robust,$
   lwindow=bartlett,lags=5) lnrxrate
*
* Tests with deviations from common time effects
*
panel(entry=1.0,time=-1.0,smpl=oecd) lnrxrate / cxrate
*
@levinlin(smpl=oecd,lags=10,crit=aic,band=5) cxrate
@htunit(smpl=oecd) cxrate
@ipshin(smpl=oecd,lags=8,crit=aic) cxrate
@breitung(smpl=oecd,det=constant) cxrate
@hadri(smpl=oecd,det=constant,variance=robust,$
  lwindow=bartlett,lags=5) cxrate
```