Paper for EML/NSF Symposium on the Bootstrap. Berkeley, July 30-Aug 6, 1996.

First Draft. July 1996

A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test: Evidence From Simulations and the Bootstrap

by

G. S. Maddala and Shaowen Wu

Department of Economics,

Ohio State University

emails: maddala.1@osu.edu and wu.139@osu.edu

Abstract

The paper compares the Levin-Lin (LL) and Im-Pesaran-Shin (IPS) panel data unit root tests with the Fisher test. The LL test has been used in several applications notably in papers on Purchasing Power Parity using panel data. This test is valid for a very restrictive hypothesis which is rarely if ever of interest in practice. The IPS test relaxes this assumption but it is argued in this paper that the Fisher test suggested by R. A. Fisher over 60 years ago is a better alternative. The paper presents some simulation evidence supporting this claim. Also, none of the tests is valid in the presence of cross correlations among the cross-sectional units. The distributions of the test statistics are unknown but the bootstrap distribution can be used to make inferences. Another alternative in this case is to use Bonferroni bounds but the Monte Carlo evidence suggests that this test is worse than the Fisher test. Thus, using the Fisher test in conjunction with the Bootstrap is the preferred alternative.

1. Introduction

Ever since the appearance of the papers by Levin and Lin (1992, 1993), panel data unit root tests have been very popular among empirical researchers having access to a panel data set. It is by now a common accepted argument that the commonly used unit root tests like the Dickey-Fuller (DF), Augmented Dickey Fuller (ADF) and Phillips-Perron (PP) tests lack power in distinguishing the unit root null against stationary alternatives, and that using panel data unit root tests is one way of increasing the power of unit root tests based on a single time series. See e.g. the arguments in Oh (1996), Wu (1996), and Frankel and Rose (1996), who try to resurrect the Purchasing Power Parity (PPP) theory using the panel data unit root tests. For a critique of these studies on PPP, see Maddala and Liu (1996).

This argument is not valid, because it is not meaningful to compare two tests that have different null hypotheses. For instance consider the simplified model

$$y_{it} = \alpha_i y_{i,t-1} + u_{it} \qquad i = 1, 2, \ldots N$$
$$t = 1, 2, \ldots T$$

Suppose we are interested in testing $\alpha_1 = 1$ vs. $\alpha_1 < 1$, we apply a single equation unit root for the first time series. The panel data unit root test, tests a different hypothesis:

$$H_o : \alpha_i = 1$$

$$\text{vs. } H_1 : \alpha_i < 1 \qquad \text{for } i = 1, 2, \ldots N$$

Furthermore, there are now more powerful tests available even in the single equation context. See e.g. Elliott et al. (1996) and also Perron and Ng(1995) who suggest modifications of the P-P test to increase power.

In the following sections, we shall present a review of the Levin-Lin tests, their extension by Im, Pesaran and Shin (1995), which will be referred to as the IPS test, and a simple alternative suggested in Maddala and Liu (1996) due to Fisher (1932) which will be referred to as the Fisher test. We then present some results from Monte Carlo experiments comparing these three tests and some evidence from the bootstrap.

2. Panel Data Unit Root Tests

Some early papers on testing for unit roots based on panel data are by Quah (1992, 1994) and Breitung and Mayor (1994). Since these have been superceded by the papers by Levin and Lin (1992, 1993), we shall not discuss them here.

2.1 The Levin-Lin (LL) test

Levin and Lin (1992) conduct an exhaustive study and develop unit root tests for the model:

$$y_{it} = \alpha y_{i,t-1} + \delta_0 + \delta_1 t + \eta_i + v_t + \varepsilon_{i,t}$$

$i = 1, 2, ..., N$

$t = 1, 2, ..., T$

Thus, the model incorporates a time trend as well as individual and time-specific effects. Initially, they assume that $\varepsilon_{it} \sim \mathrm{IID}\left(0, \sigma^2\right)$ but they say that under serial correlation with the inclusion of lagged first differences as in the ADF test the test statistics have the same limiting distributions as mentioned subsequently, provided the number of lagged differences increase with sample size. Levin and Lin consider several subcases of this model. In all cases the limiting distributions are as $N \to \infty$ and $T \to \infty$. Also in all cases, the equation is estimated by OLS as a pooled regression model. The submodels are:

Model 1: $y_{it} = \alpha y_{i,t-1} + \varepsilon_{i,t} \quad H_0: \alpha = 1$

Model 2: $y_{it} = \alpha y_{i,t-1} + \delta_0 + \varepsilon_{i,t} \quad H_0: \alpha = 1$

Model 3: $y_{it} = \alpha y_{i,t-1} + \delta_0 + \delta_1 t + \varepsilon_{i,t} \quad H_0: \alpha = 1, \delta_1 = 0$

Model 4: $y_{it} = \alpha y_{i,t-1} + v_t + \varepsilon_{i,t} \quad H_0: \alpha = 1$

Model 5: $y_{it} = \alpha y_{i,t-1} + \eta_i + \varepsilon_{i,t} \quad H_0: \alpha = 1, \eta_i = 0 \text{ for all } i$

Model 6: $y_{it} = \alpha y_{i,t-1} + \eta_{i0} + \eta_{i1} t + \varepsilon_{i,t} \quad H_0: \alpha = 1, \eta_{i1} = 0 \text{ for all } i$

For models 1-4, they show that

(a) $T\sqrt{N}(\hat{\alpha} - 1) \Rightarrow N(0,2)$

(b) $t_\alpha \Rightarrow N(0,1)$

For models 5, if $\dfrac{\sqrt{N}}{T} \to 0$, then

(a) $T\sqrt{N}(\hat{\alpha} - 1) + 3\sqrt{N} \Rightarrow N(0,10.2)$

(b) $\sqrt{1.25}t_\alpha + \sqrt{1.875N} \Rightarrow N(0, \dfrac{645}{112})$

In model 6, both intercept and time-trend vary with individuals.

In the empirical applications, Oh (1996) uses only models 1 and 5. Wu (1996) uses the complete model with trend, and individual and time specific effects but uses the distributions derived for model 5. Papell (1985) uses model 5 with lagged first differences added but computes his own exact finite sample critical values using Monte Carlo methods and finds them 3 to 15% higher than those tabulated in Levin and Lin (1992).

Levin and Lin argue that in contrast to the standard distributions of unit root test statistics for a single time series, the panel test statistics have limiting normal distributions. However, the convergence rates are faster as $T \rightarrow \infty$ (superconsistency) than as $N \rightarrow \infty$.

The paper by Levin and Lin (1993) provides some new results on panel unit root tests. However, since the models discussed are the same, we shall not go into the details. Actually, Levin and Lin suggested a four step procedure for the computation of their test statistic in the presence of heteroskedasticity and autocorrelation. We shall skip the details here but in the simulations, we used the four step procedure.

The major limitation of the Levin-Lin tests is that α is the same for all observations. Thus, if we denote by α_i the value of α for the i-th cross-section unit then the Levin-Lin test specifies the null H_0 and alternative H_1 as:

$$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_N = \alpha = 1$$

$$H_1: \alpha_1 = \alpha_2 = \ldots = \alpha_N = \alpha < 1$$

The null makes sense under some circumstances, but the alternative is too strong to be held in any interesting empirical cases. For example, in testing the convergence hypothesis in growth models, one can formulate the null as implying that none of the economies under study converges and thus $\alpha = 1$ for all countries. But it does not make any sense to assume that all the countries will converge at the same rate if they do converge.

2.2 The Im-Pesaran-Shin (IPS) Test

IPS relax the assumption that $\alpha_1 = \alpha_2 = ... = \alpha_N$ under H_1. The basic idea of the test is very simple. Take model 6 in Levin and Lin and substitute α_i for α. Essentially what we have is a model with a linear trend for each of the N cross-section units. Thus, instead of pooling the data, we use separate unit root tests for the N cross-section units. Consider the t-test for each cross-section unit based on T observations. Let t_i (i =1, 2, ..., N) denote the t-statistics for testing unit roots, and let $E(t_i) = \mu$ and $V(t_i) = \sigma^2$. Then $\sqrt{N}(\bar{t} - \mu)/\sigma \Rightarrow N(0,1)$. The problem is computing μ and σ^2. This they do by Monte Carlo methods and tabulate them for ready reference (Table A, B, C of their paper).

The important thing to note is that the IPS test is a way of combining the evidence on the unit root hypothesis from the N unit root tests performed on the N cross-section units. Note that implicit in the test is the assumption that T is the same for all cross-section units and hence $E(t_i)$ and $V(t_i)$ are the same for all i. Thus, we are considering only balanced panel data.

In the case of serial correlation, they proposed to use the ADF t-test for individual series. However, $E(t_i)$ and $V(t_i)$ will vary as the lag length included in the ADF regression varies. They tabulated $E(t_i)$ and $V(t_i)$ for different lag lengths. In practice, however, to make use of their tables, we are restricted implicitly to use the same lag length for all the ADF regressions for individual series.

2.3 Fisher's (P_λ) Test

R. A. Fisher (1932) suggested a method of combining the evidence from several independent tests. Suppose there are N unit root test as in IPS. Let P_i be the observed significance level (P-value) for the i-th test. Then ($-2\sum \log_e P_i$) has a χ^2 distribution with d. f. 2n. This is an exact distribution. This test, often known as the P_λ test, is mentioned in Rao (1952, p. 44) and Maddala (1977, Section 4.12) but has not received much attention.

The advantage of this test is that it does not require a balanced panel as in the case of the IPS test. Also, one can use different lag length in the individual ADF regression. The disadvantage is that the P-values have to be derived by Monte Carlo simulation. Another advantage is that it can also be carried out for any unit root test derived. The IPS test is easy to use because there are ready tables available in the paper for $E(t_i)$ and $V(t_i)$. However, these are valid only for the ADF test.

3. A Comparison of the Different Tests

Some broad comments on the merits and demerits of the two tests would be useful in interpreting the results to be presented later. Hence we shall go through these first.

(1) The LL tests test a very restrictive hypothesis that is rarely of practical interest.

(2) The IPS test is claimed to be a generalization of the LL tests. However, it is better viewed as a way of combining the evidence of several independent unit root tests.

(3) Im-Pesaran-Shin present power of the LL and IPS tests and argue that the IPS test is more powerful than the LL test. However, strictly speaking, the power comparison is not valid. Although the null hypothesis is the same in the two tests, the alternative hypothesis is different. The LL tests are based on homogeneity of the autoregressive parameter (although there is heterogeneity in the error variances and the serial correlation structure of the errors). Thus the tests are based on pooled regressions. The IPS test on the other hand is based on heterogeneity of the autoregressive parameter. As argued earlier, the test amounts to a combination of different independent tests. There is no pooling of data involved as in the LL tests. In the following sections, we shall also present power comparisons with the LL test but it should be borne in mind that the LL test will necessarily come out worse because the LL test has to use panel estimation method which is not efficient if there is no need for pooling.

(4) The Fisher test and the IPS test <u>are</u> directly comparable. The aim of both the tests is a combination of the significance of different <u>independent tests</u>. The Fisher test is non-parametric, whatever test statistic we use for testing for a unit root for each sample, we can get the P-values P_i and then $-2\sum \log_e P_i \sim \chi^2$ with 2N d. f., where N is the number of separate samples. The IPS test on the other hand is parametric. The

distribution of the t-bar statistic involves the mean and variance of the t-statistics used. IPS compute this for the ADF test statistic for different values of the number of lags used and different sample sizes. However, these tables are valid only if the ADF test is used for the unit root tests. Also, if the length of the time series for the different samples is different, there is a problem using the tables prepared by IPS. The Fisher test does not have any such limitations. it can be used with any unit root test and even if the ADF test is used, the choice of the lag length for each sample can be separately determined. Also, there is no restriction of the sample sizes for different samples (they can vary according to availability of the data).

(5) The Fisher test is an exact test. The IPS test is an asymptotic test. The normality of $\bar{t}$ holds only asymptotically, although IPS argue that the test is valid for small sample sizes as well.

(6) The crucial element that distinguishes the two tests is that the Fisher test is based on combining the significance levels of the different tests, and the IPS test is based on combining the test statistics, which is better is the question. We conducted Monte Carlo studies with these issues in mind.

(7) Both the Fisher test and IPS test are based on combining independent tests. So if there is contemporaneous correlation, then there are correlations among the individual test statistics. Both tests will break down in this case.

Broadly speaking, our conclusions so far (more detailed Monte Carlo experiments are being conducted) are as follows:

(1) In general, for balanced panels, when there is not much variation in the autoregressive parameter, and it is close to 1, the IPS test is slightly more powerful than the Fisher test, and both are more powerful than LL test. However, for alternatives not too close to 1 (say 0.8 or lower) the Fisher test shows higher power than IPS tests.

(2) In the case of unequal sample sizes, the Fisher test, as expected, performs better than the other two tests because it uses all the observations, whereas for the LL and IPS tests we have to throw away some observations to get a balanced panel.

(3) As for the issues of heteroskedasticity and serial correlation in the errors, all the tests can take care of these problems. But when the errors in the different samples (or cross-section units) are cross correlated (as would often be the case in empirical work) none of the tests can handle this problem well. In all the cases the correlations among the t-statistics create problems. However, the Monte Carlo evidence suggests that this problem is less severe with the Fisher test than with the LL or the IPS test. More specifically, when T is large but N is not very large, the size distortion with the Fisher test is small. But for medium values of T and large N, the size distortion of the Fisher test is of the same level as that of the IPS test.

Overall, we can conclude from these results that the Fisher test is better than the IPS and LL tests.

4. Design of the Monte Carlo Studies

In this section, we present details on the designs of Monte Carlo simulations. Basically, there are two experiments. The first one studies the size and power performances of the three unit root tests under the general setup of the null and the alternative hypotheses, i.e., $H_0: \alpha_i = \alpha = 1$ for all i and $H_1: \alpha_i = \alpha < 1$ for all i. In the second experiment, we change the alternative to $H_1: \alpha_i = 1$ for some i and $\alpha_i < 1$ for the other i. We do so due to the following consideration. In reality, some series in a panel under study might be stationary while the others are non-stationary. We expect the null to be rejected for such panels. But due to the masking problems, power might be relatively low for panel unit root tests. Thus we want to investigate the powers of the three panel unit root tests in this circumstance. Details of these two designs are as follows.

4.1 Monte Carlo simulation following Im-Peseran-Shin

In this experiment, we use the following data generating process (DGP) for a dynamic panel containing group and time specific effects:

Drift Model: $x_{it} = (1-\phi_i)\mu_i + \phi_i x_{i,t-1} + u_{it}$

Trend Model: $x_{it} = \mu_i + (1-\phi_i)\mu_i t + \phi_i x_{i,t-1} + u_{it}$

where i = 1, 2, ..., N and t = 1, 2, ..., T and x_{i0} is random.

The error term u_{it} contains time specific effect θ_t and random component ε_{it}:

$$u_{it} = \theta_t + \varepsilon_{it}$$

where $\theta_t = 0.9\theta_{t-1} + \omega_t$, $\omega_t \sim N(0,1)$ and $\varepsilon_{it} = \rho_i \varepsilon_{it-1} + e_{it}$, ρ_i's are randomly generated on U[0.2, 0.4] and different for each i where U denotes the uniform distribution. We assume that e_{it} to be jointly normal distribution with $E(e_{it}) = 0$, $E(e_{it}, e_{js}) = \begin{cases} \sigma^2 & \text{for } t = s \\ 0 & \text{for } t \neq s \end{cases}$. If we let Σ denotes $(\sigma_{ij})_{i,j=1}^{N}$, then nonzero terms on the offdiagonal terms in Σ represents the existence of cross correlations. Here we randomly generate some positive definite matrices for Σ.

In the simulation, μ_i is generated randomly from N(0,1) and then fixed for each model. The randomness of x_{i0} is achieved by generating extra x's and discarding some initial observations.

This experiment is very similar to IPS's experiments. The biggest difference is that we allow contemporaneous correlation in ε_{it}.

After the data have been generated, we apply the LL, IPS and Fisher tests. For the Fisher test, we apply ADF(p) test for each individual series

Drift Model $\quad \Delta y_{it} = \alpha_i + \beta_i y_{i,t-1} + \sum_{j=1}^{p} \gamma_{ij} \Delta y_{i,t-j} + \text{residual}$

Trend Model $\quad \Delta y_{it} = \alpha_i + \delta_i t + \beta_i y_{i,t-1} + \sum_{j=1}^{p} \gamma_{ij} \Delta y_{i,t-j} + \text{residual}$

where p = 0, 1, 2 are used and compute P-values using the Dickey-Fuller t-distributions generated by 100,000 simulations for the corresponding ADF t-test statistics. Consequently, $P_\lambda = (-2\sum \log_e P_i)$ is calculated. For the LL and IPS tests, we follow the procedures described in their papers.

In each experiment , we consider the cases of N = 10, 25, 50, 100 and t = 10, 25, 50, 100. 2000 replications is used in computing empirical size and power of the tests. In calculating the power, ϕ_I is set to be 0.9 for all i. The results are presented in Tables 1 and 2.

Table 1 reports the simulation results of the model with constant term only. For N of the medium size (for example, N = 25), the Fisher test has the smallest size distortion. As N increases, the IPS test begin to outperform the Fisher test when T is small, say T = 25. However, if T increases also, the Fisher test performs much better than the IPS test in the sense that the Fisher test has smaller size distortion with comparable power. As we have mentioned, the LL test is the worst in almost all the cases when cross correlation is taken into consideration. When reading Table 1, we should keep in mind that the direct power comparison is not appropriate since here are significant size distortions.

The argument that Fisher test is the best is still valid when a linear time trend is included in the model. The advantage is even stronger as we can see in Table 2.

Another observation in Tables 1 and 2 is the effect of selection of order of the ADF regressions. We generate the error terms according to an AR(1) model thus using ADF(1) is appropriate. From the results we can see the danger of under-selecting the order of ADF regression which is also pointed out by IPS. Meanwhile, over-selecting the order of ADF regression, i.e., p=2, seems to alleviate the size distortions for all three tests, the price is the declines in powers.

4.2 Simulation with a mixture of stationary and nonstationary alternatives

In this experiment we generate the data according to DGP described in 4.1 except that we set $\phi_i = 1$ for part of the individual series and $\phi_i < 1$ for the others. We study the case of N = 17 and T = 50 since that is the sample size of the 17 European countries data in the empirical analysis. First, we consider the case where there is only one stationary process in the 17 countries, then we add the number of stationary processes to 2, 4, 8, 10 and 12. For the 1, 2, 4 cases, we set $\phi_i = 0.8$, while for 8, 10, 12 cases, ϕ_i is generated from U[0.7, 0.9] and different for each series. The powers of the three tests are reported in Table 3. The general result is that the Fisher test has the higher power in all cases. The more the number of stationary processes included, the stronger the relative advantage. Thus if only part of the panel are stationary, the Fisher test is the most likely one to point it out.

5. The Role of the Bootstrap

Recall that the properties of all the three tests are based on the assumption that the error terms are not cross correlated. When this assumptions is violated, the derived distributions for these test statistics are no longer valid. More specifically, if there is cross-correlation in the data, as noted earlier, the distributions of the test statistics are not the same as before and are not known. For the t-bar test in IPS, the t-statistics are correlated and hence the t-bar statistic does not have the stated variance in its (asymptotic) normal distribution, and for the Fisher test, too, we do not have independent tests and hence the P_λ criterion does not have the χ^2 distribution. One way out of this is to use the bootstrap method to get the empirical distributions of the test statistics to make inferences. We shall discuss the bootstrap method briefly and then present the results.

The bootstrap method for univariate time series is well developed. Readers are referred to Li and Maddala (1996) for a good introduction. Meanwhile the bootstrap method for panel data is in its infancy. We shall discuss our proposed method in detail.

The problem here is to generate the bootstrap distributions of the test statistics. The bootstrap data were generated using the sampling scheme S_3 described in Li and Maddala (1996). Generally speaking, if the null hypothesis is $H_0: \beta = \beta_0$, the sampling scheme S_3 suggests to generate bootstrap sample y^* as $y^* = x\beta_0 + \varepsilon^*$ where ε^* is the bootstrap sample from $\varepsilon^0 = y - x\beta_0$. Since we have panel data here, we should also take care of special problems arising from the serial correlation. So we get the bootstrap sample of the error term ε^0 from

$$\Delta y_{it} = \hat{\rho}_i \Delta y_{it-1} + \varepsilon_{it}^0 \tag{5.1}$$

where $\Delta y_{it} = y_{it} - y_{it-1}$ since under the null hypothesis y_{it} has a unit root. Since there are cross correlations among ε_{it}^0, we cannot resample ε_{it}^0 directly. We proposed to resample ε_{it}^0 with cross-section index fixed, i.e., instead of resampling ε_{it}^0, we resample $\varepsilon_t^0 = [\varepsilon_{1t}^0, \varepsilon_{2t}^0, \cdots, \varepsilon_{Nt}^0]'$ to get ε_t^*. In this way, we can preserve the cross correlation structure of the error term. Then the bootstrap sample y^* is generated as

$$y_{it}^* = y_{it-1}^* + u_{it}^* \quad \text{with } y_{i0}^* = 0$$

$$u_{it}^* = \hat{\rho}_i u_{it-1}^* + \varepsilon_{it}^* \quad \text{with } u_{i0}^* = \sum_{j=0}^{m} \hat{\rho}_i^j \varepsilon_{-j}^*$$

where $\hat{\rho}_i$'s are from estimation results of (5.1). The method to get the initial value of u_{i0}^* is suggested by Rayner (1990). Here ε_{-j}^*'s are drawn as an independent bootstrap sample and m is selected as 30. 2,000 replications are used to generate the empirical distributions. They are drawn in Figures 1 to 3.

As we can see from the figures, the empirical distributions of both IPS and LL test shift to the right dramatically, which implies that when the usual critical value derived from the left tail of the standard normal is used, the IPS test and LL tests can cause under-rejection of the null hypothesis. On the other hand, the distribution for the Fisher test shifts to the left, but in a less dramatic manner, and hence it suffers from implied less severe size distortion. Since the right tail test is used for the χ^2 distribution, the Fisher test causes under-rejection as well.

We can also get the critical values at the 5% level based on these empirical distributions. For the sample of the 17 European countries under study, these critical values are -0.518, 15.80, and -1.157 for the IPS, Fisher and LL test respectively. Now, the test statistics for the same sample are -0.950, 48.912, and -1.078. So strikingly, we can strongly reject the unit root null for the data set by using the critical values at 5% level for the IPS and the Fisher test, although for the LL test, we fail marginally to reject the null

In general, the conclusions from the bootstrap experiments are:

In the case of cross-correlated errors there are substantial size distortions in using the conventional test statistics for the IPS, Fisher and LL tests, although the size distortions are less serious for the Fisher test than for the IPS test. Using the bootstrap method results in a decrease of these size distortions, although it does not eliminate them. The Fisher test does better than the IPS test overall using the bootstrap method.

6. Tests Based on Bonferroni Inequality

For the correlated errors case we argued that, for both the Fisher test statistic and the IPS test statistic, the distribution is not known but bootstrap methods could be used to make inferences. There is, however, one test statistic that is equally applicable in both the uncorrelated and correlated cases. This is the statistic based on the Bonferroni inequality (See Alt, 1982) and discussed in Dufour and Torres (1996). These tests have also been discussed earlier in Savin (1984) and Dufour (1990).

The idea behind this test is to break up the hypothesis H_0: $\alpha_i = 1$ for all i, $i = 1,2,\cdots,N$ into a set of subhypotheses $H_{0i}:\alpha_i = 1$ and noting that H_0 is wrong if and only if any of its components H_{0i} is wrong. Suppose we choose the significance level γ_i for the i-th test. Then if we follow the rule that we reject H_0 if at least one of the subhypotheses H_{0i} is rejected at significance level γ_i for H_{0i}, then the Bonferroni inequality says that the significance level γ for H_0 is given by

$$\gamma \leq \sum_{i=1}^{N} \gamma_i$$

One simple rule that Dufour and Torres suggest is to take $\gamma_i = \gamma / N$, unless there is a-priori compelling reason that some tests ought to be rejected at lower (or higher) significance level than the others.

We shall now compare this test with the Fisher test for both the cases of independent tests and correlated tests. Since the tables of significance values for unit root tests are readily available at the 1% significance level, we chose N = 5 in our investigation, so that γ_i = 0.01 and γ = 0.05. Note that the case of small N is favorable to the test based on the Bonferroni inequality (see p. 299 of Alt, 1982).

We consider 4 cases of DGP. The basic DGP is

$$y_{it} = \mu_i + \rho_i y_{i,t-1} + \varepsilon_{it}$$

$$\mu_i \sim U[0,1]$$

$$\rho_i = \begin{cases} 0.9 & \text{for power} \\ 1 & \text{for size} \end{cases}$$

where assumptions on ε_{it} are different in each case as follows:

Case 1: No serial correlation and contemporaneous correlation

$$\varepsilon_{it} \sim N(0, \sigma_i^2) \quad \text{where } \sigma_i^2 \sim U[0.5, 1.5]$$

Case 2: There is serial correlation, but no contemporaneous correlation

$\varepsilon_{it} = \lambda_i \varepsilon_{i,t-1} + e_{it}$, where $\lambda_i \sim U[0.2, 0.4]$ and $e_{it} \sim N(0, \sigma_i^2)$ with $\sigma_i^2 \sim U[0.5, 1.5]$

Case 3: There is contemporaneous correlation but no serial correlation

$$\varepsilon_t = [\varepsilon_{1t}, \varepsilon_{2t}, \cdots, \varepsilon_{Nt}]' \sim N(0, \Sigma) \quad \text{where } \Sigma = \Psi'\Psi \text{ with } \Psi_{NxN} \sim U[-1,1]$$

Case 4: There are both serial correlation and contemporaneous correlation

$$\varepsilon_{it} = \lambda_i \varepsilon_{i,t-1} + e_{it}, \quad \text{where } \lambda_i \sim U[0.2, 0.4]$$

and $e_t = [e_{1t}, e_{2t}, \cdots, e_{Nt}]' \sim N(0, \Sigma)$ where $\Sigma = \Psi'\Psi$ with $\Psi_{NxN} \sim U[-1,1]$

Table 4 presents the results from the Monte Carlo studies based on 4,000 simulations. The results clearly suggest that the power of the D-T test is substantially lower than that of the Fisher test, although there is a slight size distortion for the Fisher test in the case of correlated tests.

7. Conclusions

The paper compares the Levin-Lin and Im-Pesaran-Shin (IPS) panel data unit root tests with the Fisher test which was suggested over 60 years ago by R. A. Fisher and has a celebrated history in the statistical literature. The main conclusion of the paper is that the Fisher test is simple and straightforward to use and is a better test than the LL and IPS tests. Some other conservative tests (applicable in the case of correlated tests) based on the Bonferroni bounds have also been found to be inferior to the Fisher test.

References

Alt, F. B. (1982), Bonferroni Inequalities and Intervals. *Encyclopedia of Statistical Sciences,* Vol. 1, 294-301, eds. N. L. Johnson and S. Kotz, John Wiley & Sons. New York

Breitung, J. and W. Mayer (1994), Testing for unit roots in panel data: Are wages on different bargaining levels cointegrated? *Applied Economics*, 26, 353-361.

Dufour, J. M. (1990) Exact Tests and Confidence Sets in Linear Regressions with Autocorrelated Errors, *Econometrica*, 58, 479-494.

Dufour, J. M. and O. Torres (1996), Union-intersection and sample-split methods in econometrics with application to MA and SURE models, Paper presented at the summer meetings of the Econometric Society, Iowa City.

Elliott, G., T. J. Rothenberg and J. H. Stock (1996), Efficient tests for an autoregressive unit root, *Econometrica,*64, 813-836.

Fisher, R. A. (1933), *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh 4th Ed.

Frankel, J. A. and A. K. Rose (1996), A panel project on purchasing power parity: Mean reversion within and between countries, *Journal of International Economics*, 40, 209-224.

Levin, A. and C. F. Lin (1992), Unit root test in panel data: Asymptotic and finite sample properties. University of California at San Diego. Discussion paper #92-93.

Levin, A. and C. F. Lin (1993), Unit root test in panel data: new results. University of California at San Diego. Discussion paper #93-56.

Li, H. and G. S. Maddala (1996), Bootstrapping time series models (with discussion), *Econometric Reviews*, 15, 115-195.

Liu, P. C. and G. S. Maddala (1996), Do Panel Data Unit Root Tests Resurrect the Pruchasing Power Parity Theory?, Manuscript, Ohio State University.

Maddala, G. S. (1977), Econometrics. McGraw-Hill, New York.

Maddala, G. S. and P. C. Liu (1996), Panel data unit root tests: What do they test?, Paper prepared for the International Conference on Panel Data, Amsterdam, June 1996.

Maddala, G. S. and S. Wu (1996), A Bayesian analysis of cross-country growth regressions. Paper presented at the NBER-NSF meeting on Bayesian Inference in Econometrics and Statistics, Chicago, May 1996.

Oh, K. Y. (1996), Purchasing power parity and unit root tests using panel data. *Journal of International Money and Finance*, 15, 405-418.

Papell, D. H. (1995), Searching for stationarity: Purchasing power parity under the current float, Dept. of Economics, University of Houston.

Perron, P. and S. Ng (1995), Useful modifications to some unit root tests with dependent errors and their local asymptotic properties. Forthcoming in *Review of Economic Studies.*

Quah, K. (1992), International patterns of growth I: Persistence in cross-country disparities. LSE Working Paper.

Quah, D. (1994), Exploiting cross-section variation for unit root inference in dynamic data. *Economics Letters*, 44, 9-19.

Rao, C. R. (1952) *Advanced Statistical Methods in Biometric Research*. (Wiley, New York)

Rayner, R. K. (1990), Bootstrapping p values and power in the first-order autoregression: A Monte Carlo investigation, *Journal of Business & Economic Statistics*, 8, 251-263.

Savin, N. E. (1984), Multiple Hypothsis Testing, *Handbook of Econometrics*, eds. Z. Griliches and M. D. Intriligator, Ch. 14, North-Holland, Amsterdam.

Wu, Y. (1996), Are real exchange rates stationary? Evidence from a panel data test. *Journal of Money, Credit and Banking*, Vol. 28, No.1, 54-63.

Table 1 Size and Power of the Unit Root Tests (constant term only)

N	T		ADF(0)		ADF(1)		ADF(2)	
			Size	Power	Size	Power	Size	Power
25	25	IPS	0.0065	0.0095	0.1060	0.4235	0.0590	0.2535
		Fisher	0.0060	0.0005	0.0440	0.1910	0.0250	0.0875
		LL	0.0230	0.0245	0.1585	0.4760	0.0880	0.2575
	50	IPS	0.0025	0.0650	0.0880	0.9000	0.0645	0.8085
		Fisher	0.0075	0.0175	0.0845	0.7575	0.0665	0.6300
		LL	0.0165	0.0205	0.1415	0.7660	0.1030	0.5550
	100	IPS	0.0010	0.7725	0.1050	1.0000	0.0920	0.9995
		Fisher	0.0015	0.1090	0.0230	0.9975	0.0175	0.9925
		LL	0.0220	0.0265	0.1440	0.9905	0.1300	0.9525
50	25	IPS	0.0000	0.0060	0.1040	0.6520	0.0400	0.3705
		Fisher	0.0040	0.0025	0.1070	0.4945	0.0575	0.2695
		LL	0.0115	0.0170	0.1805	0.6950	0.0840	0.3555
	50	IPS	0.0015	0.0690	0.1035	0.9920	0.0715	0.9620
		Fisher	0.0025	0.0040	0.0730	0.9355	0.0510	0.8250
		LL	0.0080	0.0105	0.1550	0.9400	0.1030	0.7865
	100	IPS	0.0000	0.9880	0.0905	1.0000	0.0745	1.0000
		Fisher	0.0005	0.6970	0.0545	1.0000	0.0410	1.0000
		LL	0.0025	0.0800	0.1410	1.0000	0.1130	0.9990
100	25	IPS	0.0000	0.0030	0.0985	0.8615	0.0235	0.5415
		Fisher	0.0000	0.0005	0.1315	0.7080	0.0495	0.4040
		LL	0.0040	0.0070	0.1950	0.8760	0.0660	0.5085
	50	IPS	0.0000	0.0830	0.0945	1.0000	0.545	0.9995
		Fisher	0.0005	0.0025	0.0740	0.9970	0.0415	0.9780
		LL	0.0010	0.0075	0.1770	0.9990	0.0990	0.9495
	100	IPS	0.0000	1.0000	0.0875	1.0000	0.0645	1.0000
		Fisher	0.0000	0.9010	0.0460	1.0000	0.0345	1.0000
		LL	0.0005	0.0790	0.1575	1.0000	0.1225	1.0000

Table 2 Size and Power of the Unit Root Tests (time trend)

N	T		ADF(0)		ADF(1)		ADF(2)	
			Size	Power	Size	Power	Size	Power
25	25	IPS	0.0010	0.0005	0.1145	0.2000	0.0450	0.0900
		Fisher	0.0010	0.0000	0.1060	0.1805	0.0535	0.0765
		LL	0.0015	0.0020	0.1090	0.2285	0.0150	0.0320
	50	IPS	0.0000	0.0015	0.1055	0.4895	0.0655	0.3340
		Fisher	0.0000	0.0000	0.0940	0.3670	0.0525	0.2305
		LL	0.0000	0.0030	0.1180	0.5635	0.0365	0.2430
	100	IPS	0.0000	0.0190	0.1015	0.9895	0.0780	0.9620
		Fisher	0.0000	0.0040	0.0435	0.9210	0.0335	0.8390
		LL	0.0000	0.0130	0.1250	0.9665	0.0650	0.8420
50	25	IPS	0.0000	0.0000	0.1320	0.2635	0.0365	0.0790
		Fisher	0.0000	0.0000	0.1340	0.2295	0.0465	0.0780
		LL	0.0000	0.0005	0.1135	0.2850	0.0065	0.0195
	50	IPS	0.0000	0.0000	0.1155	0.6920	0.0705	0.4775
		Fisher	0.0000	0.0000	0.0705	0.4560	0.0385	0.2625
		LL	0.0000	0.0010	0.1165	0.7420	0.0235	0.3170
	100	IPS	0.0000	0.0275	0.0975	1.0000	0.0720	0.9990
		Fisher	0.0000	0.0005	0.0420	0.9970	0.0285	0.9830
		LL	0.0000	0.0100	0.1210	0.9990	0.0530	0.9710
100	25	IPS	0.0000	0.0000	0.1265	0.3305	0.0185	0.0700
		Fisher	0.0000	0.0000	0.1570	0.2850	0.0360	0.0785
		LL	0.0000	0.0000	0.0925	0.3335	0.0010	0.0085
	50	IPS	0.0000	0.0000	0.1105	0.9015	0.0470	0.6930
		Fisher	0.0000	0.0000	0.0645	0.7100	0.0260	0.4215
		LL	0.0000	0.0000	0.1155	0.9220	0.0130	0.4525
	100	IPS	0.0000	0.0240	0.0900	1.0000	0.0645	1.0000
		Fisher	0.0000	0.0005	0.0355	1.0000	0.0265	0.9995
		LL	0.0000	0.0030	0.1255	1.0000	0.0440	0.9990

Table 3 Power Comparison: Design 2 (N = 25, T= 50)*

Number of Stationary Units	IPS	Fisher	LL
1	0.0785	0.1085	0.0885
2	0.1270	0.1875	0.1040
4	0.1835	0.2490	0.1350
8	0.4685	0.5900	0.2545
10	0.6165	0.6840	0.3240
12	0.7385	0.8145	0.4045

*In the Monte Carlo simulation design, for the cases of stationary units included is 1, 2, 4, 8, we set AR(1) coefficients as 0.8. In the cases of 10 and 12, AR(1) coefficients are generated at U[0.7, 0.9].

Table 4 Size and Power comparison of the Fisher Test With the Dufour-Torres(DT) Test

Case #	Test	Size	Power
Case 1 T = 25	Fisher DT	.0505 .0460	.1180 .0693
Case 1 T = 50	Fisher DT	0.473 .0467	.3142 .1232
Case 2 T = 25	Fisher DT	.0560 .0565	.1135 .0760
Case 2 T = 50	Fisher DT	.0480 .0493	.2802 .1185
Case 3 T = 25	Fisher DT	.0785 .0480	.1630 .0683
Case 3 T = 50	Fisher DT	.0750 .0432	.3448 .1215
Case 4 T = 25	Fisher DT	.0858 .0578	.1530 .0767
Case 4 T = 50	Fisher DT	.0825 .0462	.3073 .1143