MPRA
Munich Personal RePEc Archive

Unit Roots in Economic and Financial Time Series: A Re-Evaluation based on Enlightened Judgement

Kim, Jae and Choi, In

17 December 2015

Online at https://mpra.ub.uni-muenchen.de/68411/
MPRA Paper No. 68411, posted 18 Dec 2015 00:37 UTC

Unit Roots in Economic and Financial Time Series: A Re-Evaluation based on Enlightened Judgement

Jae H. Kim*
Department of Economics and Finance
La Trobe University, VIC 3086
Australia

In Choi
Department of Economics
Sogang University
Seoul, Korea

Abstract

This paper re-evaluates the key past results of unit root test, emphasizing that the use of a conventional level of significance is not in general optimal due to the test having low power. The optimal levels for popular unit root tests, chosen using the line of enlightened judgement under a symmetric loss function, are found to be much higher than conventional ones. We also propose simple calibration rules for the optimal level of significance for a range of unit root tests based on asymptotic local power. At the optimal levels, many time series in the extended Nelson-Plosser data set are judged to be trend-stationary, including real income variables, employment variables and money stock. We also find nearly all real exchange rates covered in the Elliott-Pesavento study to be stationary at the optimal levels, which lends strong support for the purchasing power parity. Additionally, most of the real interest rates covered in the Rapach-Weber study are found to be stationary.

JEL codes: C12, E30, F30

Keywords: Expected Loss; Optimal Level of Significance; Power of the Test; Response Surface

December 2015

* Corresponding author: Tel: +613 9479 6616; Email address: J.Kim@latrobe.edu.au

All computations are conducted using R (R Core Team, 2014) and its package urca (Pfaff, 2008). The R codes used for this study are available on request. Constructive comments from Amelie Charles, Olivier Darne, Graham Elliott, Tom Engsted, Lutz Kilian, Tae Hwy Lee, Abul Shamsuddin, Dick Startz, Tom Stanley, and Herman van Dijk are gratefully acknowledged.

1. Introduction

Since the seminal study by Nelson and Plosser (1982), the presence of a unit root in economic and financial time series has been a highly controversial issue. It has compelling implications for a wide range of economic and financial theories. For example, a unit root in the real GNP contradicts the conventional view of the business cycle that a shock to the economy has a transitory effect (see Campbell and Mankiw; 1987). However, a series of studies report mixed and inconclusive results regarding the presence of a unit root in the U.S. real GNP (see, for example, Rudebusch, 1993; Diebold and Senhadji, 1996; Murray and Nelson, 2000; Papell and Prodan, 2004; Darné, 2009; Luo and Startz, 2014). Other research areas where the presence of a unit root is contentious include empirical studies on the purchasing power parity (Lothian and Taylor, 1996; Papell, 1997); and the stationarity of real interest rate (Rose, 1988; Rapach and Weber, 2004). See Choi (2015) for further discussions on economic issues related to the presence of a unit root.

A major problem of unit root testing is that the power of the test is seriously low in small samples (see, McCallum, 1986, p.406; Dejong et al., 1992; Schwert, 1992). However, the low power is not fully taken into account in its practical implementations (Cochrane, 1991, p.283). Specifically, the test is almost exclusively conducted at the conventional level of significance (typically at 0.05), completely ignoring its power and other factors. To this end, a number of authors have raised serious concerns and criticisms about the fact that empirical researchers pay little attention to the power or probability under the alternative hypothesis (Hausman, 1978; MacKinnon, 2002, p.633; Ziliak and McCloskey; 2008; Startz, 2014). It also has been pointed out that employing a conventional level is arbitrary and can lead to misleading results (Keuzenkamp and Magnus, 1995; Davidson and MacKinnon, 1993, p.79; Lehmann and

Romano, 2005, p.57). In the context of unit root testing, Maddala and Kim (1998, p.128) question the appropriateness of using the conventional level.

It has been argued that when the power of the test is low, the level of significance should be chosen at a much higher level than 0.05 (see, for example, Kish, 1959[1]). More specifically, Winer (1962) states "when the power of the tests is likely to be low …, and when Type I and Type II errors are of approximately equal importance, the 0.3 and 0.2 levels of significance may be more appropriate than the .05 and .01 levels" (cited in Skipper et al., 1967)[2]. This will achieve a sensible balance between the probabilities of Type I and II errors, and make it possible to conduct the test with a higher power. Stressing that the level of significance should be set as a decreasing function of sample size, Leamer (1978, Chapter 4) shows how the optimal level of significance can be chosen in explicit consideration of the power and losses from wrong decisions, based on what he calls the line of enlightened judgement. Recently, Kim and Ji (2015) demonstrate how Leamer's (1978) method can be applied to empirical research in finance for more credible significance testing.

The purpose of this paper is to re-evaluate the key past results of unit root testing at the optimal level of significance chosen in explicit consideration of the power and expected loss, following Leamer (1978). It is found that the optimal levels for popular unit root tests, such as the augmented Dickey-Fuller (Dickey and Fuller, 1979; ADF) and DF-GLS tests (Elliott et al., 1996) are much higher than 0.05. In fact, they are in the 0.2 to 0.4 range for the sample sizes widely used in practice, under a symmetric loss function and equal chance for the null and alternative hypotheses. Through extensive simulations, we obtain simple calibration rules for

[1] Reprinted in Morrison and Henkel (1970, p.139).
[2] Reprinted in Morrison and Henkel (1970, p.157).

the optimal levels of significance for the ADF, Philips-Perron (1988), DF-GLS, and point optimal (ERS-P) tests of Elliott et al. (1996). When the ADF and DF-GLS tests are conducted at the optimal levels, many time series in the Nelson-Plosser data set are found be trend-stationary including the real income and money stock. For the real exchange rates examined by Elliott and Pesavento (2006), the ADF and DF-GLS tests conducted at the optimal level favor the stationarity for nearly all rates, generating strong support for the purchasing power parity. Furthermore, most of the real interest rates series covered in Rapach and Weber's (2004) study are found to be stationary at the optimal levels. We also demonstrate how the calibration rules for the optimal levels of significance for the Phillips-Perron and ERS-P tests are used to determine the presence of a unit root in the U.S. real GNP. This paper is organized as follows: Section 2 presents the line of enlightened judgement and the optimal levels of significance for the ADF and DF-GLS tests; Section 3 presents the calibration rules based on asymptotic local power for a range of popular unit root tests; Section 4 re-evaluates the past key results of unit root testing; and Section 5 concludes the paper.

2. Optimal Level of Significance for Unit Root Tests

In this section, we introduce Leamer's (1978) line of enlightened judgement for the ADF and DF-GLS tests and present their optimal levels of significance under a range of sample sizes widely encountered in practice. We also examine the effects of other factors (prior probability, relative loss, starting values of the series) that can influence the choice of the optimal level.

2.1 Line of Enlightened Judgement and Optimal Level of Significance

Let α represent the level of significance which is the probability of rejecting the true null hypothesis (Type I error). The probability of Type II error (accepting the false null hypothesis) is denoted as β, with $(1-\beta)$ being the power of the test. A trade-off between α and β is well-known, with a higher (lower) value of α associated with a lower (higher) value of β. When α is set at 0.05, a low power means that the value of β is much higher than 0.05. For example, if the power is as low as 0.20, there is a serious imbalance between α and β, with the latter being 16 times higher than the former. As a result, the test is severely biased towards Type II error, with a consequence that a false null hypothesis is frequently accepted. By choosing a higher value of α in this case, say 0.3, one can achieve a sensible balance between α and β, obtaining a higher power at the same time. The line of enlightened judgement (Leamer, 1978) is formulated by plotting the combination of all possible α and β values, from which one can choose a desired combination in explicit consideration of the power and losses under Type I and II errors.

In what follows, the line of enlightened judgement for unit root tests is presented. Following De Jong et al. (1992), we consider the time series

$$Y_t = \gamma_0 + \gamma_1 t + X_t;\ X_t = \tau X_{t-1} + u_t,\ (t = 1, \ldots, n.) \tag{1}$$

where u_t is a stationary time series with zero mean and fixed variance σ^2. The standardized initial value of (1) is denoted as $X_0^* \equiv X_0 / \sigma = (Y_0 - \gamma_0)/\sigma$. The model (1) can be re-written in the ADF form as

$$\Delta Y_t = \delta_0 + \delta_1 t + \lambda Y_{t-1} + \sum_{j=1}^{m} \rho_j \Delta Y_{t-j} + e_t\,. \tag{2}$$

Taking the ADF test as an example, the test statistic for H_0: $\lambda=0$; H_1: $\lambda < 0$ is $ADF = \hat{\lambda} / se(\hat{\lambda})$ where $\lambda \equiv (\tau-1)$ and $\hat{\lambda}$ is the least-squares (LS) estimator for λ and se($\hat{\lambda}$) is its standard error estimator. Note that $\alpha = P(ADF < CR(\alpha) \mid \lambda = 0)$ where $CR(\alpha)$ is the α-level critical value; and $\beta = P(ADF > CR(\alpha) \mid \lambda < 0)$. The line of enlightened judgement is obtained by plotting all possible combinations of (α_i,β_i), where $CR(\alpha_i)$ is the critical value corresponding to α_i.

A Monte Carlo experiment with the number of trials 10000 is conducted to calculate the (α, β) values, using MacKinnon's (1996) critical values. Following De Jong et al. (1992), the data is generated from model (2) with m =1, ρ_1 = 0.5, $\gamma = \delta = 0$, setting $\lambda = \lambda_1$ where λ_1 is a value of λ under H_1. According to De Jong et al. (1992), $\lambda_1 \in$ [-0.15, 0] is a plausible range of the parameter values under H_1. In particular, they recommend λ_1 = -0.15 for annual time series; and λ_1 = -0.05 and -0.01 for quarterly and monthly data respectively. From a grid of the α values between 0.01 and 0.99 with an increment of 0.02, the proportion of Type II error is obtained as an estimate of β. The standardized initial value X_0^* is set at 1.5, which is the most plausible value according to De Jong et al. (1992).

We also present the line of enlightened judgement for the DF-GLS test of Elliott et al. (1996), which is well-known to have a higher power when the initial variable is small. The test involves a simple modification to the ADF test by de-trending the deterministic component for more efficient estimation. For the ADF tests and DF-GLS test (for the model with a constant only), the lines of enlightened judgement are constructed using appropriate MacKinnon's (1996) critical values. For the model with a constant and linear trend, the DF-GLS test statistic follows a limiting distribution different from that of the ADF. For this case, we use the critical values from the asymptotic distribution of the test statistic obtained by simulation following Elliott et al. (1996)[3].

[3] Cheung and Lai (1995) provide response surface estimates for the critical values of the DF-GLS test for the model with a linear trend. However, they are only applicable for 5% and 10% levels of significance.

According to Leamer (1978), the expected loss from hypothesis testing is $p\alpha L_1$ + $(1\text{-}p)\beta L_2$, where $p \equiv P(\mathrm{H}_0)$, L_1 represents the loss of Type I error and L_2 that from the Type II error. Given the combinations of (α, β) values on the line of enlightened judgement, the optimal level of significance α* can be chosen so that the expected loss is minimized. The value of Type II error probability corresponding to α* is denoted as β*. The specific values of p, L_1 and L_2 depend on contexts and the researcher's prior belief. For the purpose of simplicity, initially assume that $p = 0.5$ and $L_1 = L_2$, where the minimization of the expected loss is simplified to that of $\alpha+\beta$. These assumptions mean that the researcher gives an equal weight to the two states of nature (H_0 and H_1) with a prior belief that: firstly, they are equally likely to be true; and secondly, the losses from Type I and II errors are identical. In the analysis that follows, different values of p and L's will be considered.

Figure 1 presents the lines of enlightened judgement for the ADF and DF-GLS tests for the model with constant and linear trend, when $\lambda_1 = -0.15$ under the sample sizes ranging from 60 to 130. These settings are suitable for annual time series. The line shifts towards the origin as the sample size increases, corresponding to lower values of β (or higher power) for a given value of α. The blue square dots represent the points of (α^*,β^*) where the expected loss ($\alpha+\beta$) is minimized. The optimal levels of the DF-GLS test are much lower than the ADF test due to its higher power. For all sample sizes the optimal levels are in the neighborhood of 0.3, except when n = 60 for the ADF test, which is consistent with Winer's (1962) assertion. They also decrease with sample size as Leamer (1978) suggests. From Figure 1, when n = 100 and α = 0.05, β = 0.74 and the power of the ADF test is only 0.26: i.e. as mentioned above, a case of low power with a clear imbalance between α and β. However, if α is chosen to minimize the expected loss ($\alpha+\beta$), (α*,β*) = (0.31, 0.22) with a substantially higher power of 0.78. That is, a reasonable balance between the two error probabilities is reached, and the power is nearly three times higher than when $\alpha = 0.05$. The expected loss is much higher when $\alpha = 0.05$, as expected. The critical value at the optimal level of significance is -2.54, which is much larger than the 5% critical value of -3.46. Similar results are evident for the DF-GLS test.

Figure 2 presents the lines of enlightened judgement associated with the ADF and DF-GLS tests for the model with constant only, when $\lambda_1 = -0.05$ for the sample sizes ranging from 80 to 240. These settings are suitable for quarterly time series. Again, higher power associated with the DF-GLS test is clear with the lines for the DF-GLS test much closer to the origin than those of the ADF. When the sample size is 120 and $\alpha = 0.05$, the DF-GLS test is again severely biased towards the Type II error, with its β value more than 11 times higher than that of α. The power is 0.44, which is much higher than that of the ADF which is 0.13, as expected. However, at the optimal level of significance $\alpha^* = 0.23$, the DF-GLS test enjoys a substantially higher power of 0.92 with a sensible balance between the two error probabilities. Overall, the optimal levels of significance for the DF-GLS test are in the neighborhood of 0.20 for a typical quarterly time series.

From Figures 1 and 2, we observe a tendency where the optimal level (α^*) is more or less twice the size of the corresponding Type II error probability (β^*). This means that the test at the optimal level is mildly more conservative for the Type II error, in contrast with the case of $\alpha =$ 0.05 where the test is severely biased towards Type II error. It should be noted that the conventional levels of significance (such as 0.05) represent a poor benchmark level for these tests because they cannot be optimal under any sample sizes frequently encountered in practice.

2.2 Factors Affecting the Optimal Level of Significance

Koop and Steel (1994, p.99) consider the lack of formal development of loss function as a serious weakness of both Bayesian and classical unit root studies. They argue that the classical analysis has an implicitly defined loss function in choosing the level of significance, in which losses are asymmetric. That is, the use of a conventional level of significance (such as 0.05) implies a loss function, which is arbitrarily asymmetric. While our analysis so far assumes a symmetric loss function ($L_1 = L_2$), it is possible that the value of the optimal level changes in response to different values of relative loss from Type I and II errors[4]. In addition, there are other factors that possibly affect the optimal level; i.e., the probability for the null hypothesis (p) that is so far assumed to be 0.5; and the starting value of the series that may affect the power of a unit root test.

To examine the effects of the prior probability for H_0 and the relative loss, Figure 3 plots the optimal level of significance for the ADF and DF-GLS tests (model with constant only) as a function of p and relative loss. Letting $k = L_2/L_1$ (relative loss) and setting $L_1 = 1$ without loss of generality, the expected loss is expressed as $p\alpha + (1\text{-}p)\beta k$. The optimal values are calculated from the lines of enlightened judgement given in Figure 2 when $n = 120$, by minimizing the expected loss $p\alpha + (1\text{-}p)\beta k$ under different values of p and k. It appears that the optimal level changes sensitively to the value of p and k, and that the conventional levels of significance (such as 0.05 and 0.01) are justifiable only when p is high and k is low. That is, either when the researcher has a strong prior belief that H_0 is true (presence of a unit root) or when the loss from Type I error considerably outweighs the loss associated with Type II error. In the opposite

[4] It is possible that the analysis in this paper can be conducted using a particular form of loss function as in Koop and Steel (1994). However, according to Koop and Steel (1994), the choice of loss function depends on contexts and the nature of the empirical analysis. Due to this difficulty, we consider the optimal level under different values of relative loss.

case, the optimal level can often be far higher than 0.50 for both tests. Under moderate values of p and k, the optimal levels are in the 0.2 to 0.4 range.

It is well-known that the power of unit root test changes sensitively to the initial value and the degree of autocorrelation on the error term (see Dejong et al., 1992; Müller and Elliot, 2003). To examine the sensitivity, the optimal levels of significance for the ADF and DF-GLS tests (model with a constant only) are reported in Table 1 when n = 120, under a range of X_0^* and ρ_1. For the ADF test, under a reasonable value of X_0^* (0 to 5), it appears that the optimal level of significance is not sensitive to ρ_1. For the DF-GLS test, the optimal level changes sensitively, especially when the value of ρ_1 is negative, and tends to increase with the starting value. The optimal level's sensitivity to the values of p and k; and starting value of the series are taken into account in the calibration rules, which are discussed in the following section.

3. Calibration Rules based on Asymptotic Local Power

In the previous section, we demonstrate how the optimal level of significance can be chosen in small samples. However, the choice depends on a range of factors such as the sample size, the value of λ_1, the degree of autocorrelation, and data frequency. We also observe that the prior probability of the null hypothesis (p), relative loss from Type I and II errors (k), and starting values of the series (X_0^*) play their roles. To simplify the choice in practice, it is useful to consider the asymptotic local power of a unit root test, which depends largely on the local-to-unity coefficient. By doing this, we do not need to choose the particular values of sample size and λ_1, also isolating the effects of nuisance parameters that are asymptotically negligible, such as the degree of autocorrelation. Building on this idea, we develop a simple calibration rule for the optimal level of significance of a unit root test, which uses the value of local-to-unity coefficient as a key input.

To achieve this, we follow Elliott et al. (1996) to generate the asymptotic local power as a function of local-to-unity coefficient $c \equiv n(1-\tau)$. Figure 4 presents the lines of enlightened judgement for the ADF and DF-GLS tests (the model with a constant only) under a selected values of c when $n = 500$. All other computational details are the same as before. As might be expected, the optimal level of significance α^* is a decreasing function of c. This is because the test gains a higher power as the model moves away from the unit root. To obtain the calibration rules, we calculate the asymptotic power (and the values of β) for a grid of c values ranging from 0.1 to 30 with an increment of 0.6, under different values of p and k values used in Figure 3; and X_0^* values ranging from 0 to 5. For all combinations, the optimal level is chosen so that the expected loss $p\alpha + (1-p)\beta k$ is minimized. In addition to the ADF and DF-GLS tests, we obtain the calibration rules for the Phillips-Perron and ERS-P tests, which are also widely used in practice. The calibration rules are summarized in the following response surface estimates:

- Model with a constant only

ADF: $\hat{\alpha}^* = 0.825 - 0.505p + 0.028k - 0.025c - 0.002X_0^*$

Phillips-Perron: $\hat{\alpha}^* = 0.827 - 0.516p + 0.029k - 0.025c - 0.002X_0^*$

DF-GLS: $\hat{\alpha}^* = 0.525 - 0.339p + 0.018k - 0.019c + 0.025X_0^*$

ERS-P: $\hat{\alpha}^* = 0.509 - 0.309p + 0.017k - 0.018c + 0.026X_0^*$

- Model with a constant and a linear trend

ADF: $\hat{\alpha}^* = 0.933 - 0.655p + 0.037k - 0.023c - 0.001X_0^*$

Phillips-Perron: $\hat{\alpha}^* = 0.932 - 0.665p + 0.037k - 0.023c - 0.001X_0^*$

DF-GLS: $\hat{\alpha}^* = 0.802 - 0.546p + 0.031k - 0.024c + 0.017X_0^*$

ERS-P: $\hat{\alpha}^* = 0.793 - 0.541p + 0.031k - 0.023c + 0.018X_0^*$

These response surfaces indicate that the optimal level of significance is related negatively to the distance of the model to unit-root non-stationarity (c) and $p \equiv P(H_0)$, while it is positively related to the relative loss k. Referring to the starting values, the optimal levels of the ADF and Phillips-Perron tests show negative responses but these have fairly small effects, while those of the DF-GLS and ERS-P tests show positive and large responses.

Consider a researcher who has a sample of size 120 and possible τ value of 0.95 (c = 6). Suppose this researcher is neutral about the probability of H_0 and the relative loss (p=0.5; k=1), and the starting value takes a moderate value of 1.5. Then, for the ADF test with a constant, the expected value of optimal level of significance is 0.49; and for the DF-GLS test it is 0.30. These values are close to the exact values reported in Figure 2. The above calibration rules can also be used to assess the values of p or k, implied by a preferred level of significance. For example, if one wishes to maintain the 0.05 level of significance, she can obtain the value of p which satisfies the above response surface given the values of k and X_0^*. In Section 4.4 we demonstrate with an example how these calibration rules can be used in practice.

4. Re-evaluation of Past Empirical Results

In this section, we apply the optimal levels of significance and the calibration rules to the key unit root testing results reported in previous studies. We examine the extended Nelson-Plosser data set for U.S. macroeconomic time series, the real exchange rates covered by Elliott and Pesavento (2006), and the real interest rates studied by Rapach and Weber (2004).

4.1 Extended Nelson-Plosser Data

Table 2 reports the results for the extended Nelson-Plosser data. With the ADF test, every time series, except for the real GNP, real per capita GNP, and unemployment rate, is found to have a unit root. The real GNP and real per capita GNP have their *p*-values close to 0.05, which leads to accepting the null hypothesis at the 1% level of significance. However, if the optimal level of significance (α* = 0.37) is utilized, the presence of unit roots in the real GNP and real per capita GNP is clearly rejected. At the optimal levels, the employment and money stock series are also found to be trend-stationary, in contrast to the outcomes at the conventional level. Similar results are evident when the DF-GLS test is used. That is, the unit root hypotheses for the real GNP and real per capita GNP are rejected at the optimal level of significance (α* = 0.25); and so are those for the employment and money stock series (α* = 0.21). For all other time series, the inferential outcomes of the ADF and DF-GLS tests are consistent at the conventional and optimal levels of significance.

It is interesting to observe that the above results largely agree with the Bayesian results of Schotman and van Dijik (1991). At the optimal levels, eight time series are found to be trend-stationary, namely the nominal GNP, GNP deflator, consumer prices, wages, real wages, velocity, interest rate, and common stock prices. Schotman and van Dijik (1991) find these eight to have a unit root based on the Bayesian method (with posterior probability higher than 0.75). Overall, similar to Schotman and van Dijik (1991), we find that the real variables are found to be trend-stationary while the nominal ones are difference-stationary.

4.2 Elliott-Pesavento Data

Elliott and Pesavento (2006) examine the stationarity of fifteen currencies' real exchange rates using quarterly data from 1973 to 2003. Concerned with the low power of unit root tests, they attempt to improve the power by considering the co-variation of the real exchange rates with other relevant economic variables. In so doing, they report the results of the ADF and DF-GLS tests, which accept the null hypothesis of unit root for most of the real exchange rates at the 5% level of significance. Table 3 presents the ADF and DF-GLS statistics, along with the inferential outcomes based on the optimal level of significance obtained in Figure 2. For the ADF test, at the 5% level, all exchange rates are found to have a unit root with all statistics greater than the critical value of -2.89. For the ERS-GRS test at the 5% level, all rates except for those of Belgium, Denmark, France, Germany, Italy, Norway, and Sweden are found to have a unit root. Hence, at the 5% level of significance, the purchasing power parity is not strongly supported. However, at the optimal levels obtained in Figure 2, the ADF and DF-GLS test statistics are less than the corresponding critical values rejecting the unit root hypothesis for all real exchange rates, except for the Canadian rate. These results represent strong empirical support for the validity of purchasing power parity. We note that the two tests clearly agree when the optimal levels of significance are employed.

4.3 Rapach-Weber Data

Rapach and Weber (2004) employ a range of unit root tests to examine the stationarity of real interest rates of a number of international capital markets. Using quarterly data from 1957 to 2000, they report the ADF and DF-GLS test results for 10 capital markets. The results are reported in Table 4. If the ADF test is used, the presence of unit root cannot be rejected at the

5% level for all rates except for those of Denmark and the UK. Hence, the results are strongly in favor of the presence of unit root in real interest rates. At the optimal level of significance the results are largely reversed. That is, the unit root hypothesis is rejected for all rates except for those of the Netherlands and New Zealand, providing the evidence that eight out ten rates are stationary. With the DF-GLS test, the unit root hypothesis is rejected for four of ten real interest rates, at the 5% level of significance. At the optimal levels, the DF-GLS test rejects the null hypothesis for two additional rates (Canada and Italy), providing inferential outcome that six of ten rates are stationary. Hence, at the optimal levels of significance, both unit root tests favour the stationarity of real interest rates.

4.4 Application of the Calibration Rules

In this sub-section, we demonstrate how the calibration rules can be used in practice. We take the real GNP from the extended Nelson-Plosser data set as an example, and employ the calibration rules for the Phillips-Perron and ERS-P tests. For the real GNP in natural log (denoted *Y*), LS estimation of an AR(2) model with a constant and a linear time trend provides the following results:

$$Y_t = 0.81 + 0.006\, t + 1.23\, Y_{t-1} - 0.41\, Y_{t-2}.$$

The Phillips-Perron statistic is -2.83 with the p-value of 0.19, indicating the acceptance of the unit root hypothesis at the 5% level of significance. The ERS-P test statistic is 5.64, leading to the acceptance of the null hypothesis of a unit root at the 1% level with its p-value slightly less than 0.05. Hence, at the conventional levels of significance, these two tests provide evidence that favors the presence of a unit root in the real GNP.

In obtaining the optimal level of significance from the response surface estimates, we assume that the researcher is impartial between H_0 and H_1 under a symmetric loss function, namely p

= 0.5 and k = 1. We estimate the local-to-unity coefficient as $\hat{c} \equiv n(1-\hat{\tau})$ where $\hat{\tau}$ is an estimator for AR(1) coefficient. It is well known that the parameter cannot be estimated consistently. For an AR(p) model with $p > 1$, the sum of AR coefficient estimators can be used as an estimator for $\hat{\tau}$. One may use the LS estimator, but it is well-known to be biased in small samples under-estimating the value of τ, which may result in over-estimation of c. Due to this problem, we propose the use of Kim's (2004) bias-corrected estimators for AR(p) parameters unbiased to order n^{-1}, which is a generalized version of the bias-corrected estimator for the AR(1) model of Orcutt and Winocur (1969). Kim's (2004) method makes use of the asymptotic bias formulae derived by Stine and Shaman (1989), and employs Kilian's (1998) stationarity-correction in the event that bias-correction pushes the model to non-stationarity[5]. The bias-corrected estimation gives

$$Y_t = 0.62 + 0.004\, t + 1.27\, Y_{t-1} - 0.40\, Y_{t-2},$$

and the resulting estimate for c is 10.62. This value is substantially smaller than the estimate of c based on the LS estimator, which is 14.11. The estimate of the standardized starting value X_0^* is 2.77. Plugging these c and X_0^* values (along with p=0.5, k=1) to the response surface for the Phillips-Perron test

$$\hat{\alpha}^* = 0.932 - 0.665p + 0.037k - 0.023c - 0.001X_0^*,$$

the estimated value of the optimal level is 0.39. For the ERS-P test with response surface estimate

$$\hat{\alpha}^* = 0.793 - 0.541p + 0.031k - 0.023c + 0.018X_0^*,$$

the optimal level estimate is 0.36. Hence, at the optimal levels of significance, the Phillips-Perron and ERS-P tests convincingly reject the unit root hypothesis for the real GNP. These

[5] The R package BootPR (Kim, 2015) provides computational resources for this bias-corrected estimation.

results are consistent with those in Section 4.1 where the ADF and DF-DLS tests are conducted at the optimal level.

Is should be noted that the above analysis is based on the assumption that $p = 0.5$ and $k = 1$. While these values may represent the most neutral and impartial state of nature, other values may be chosen. As we have shown previously, the value of optimal level changes under different values of p and k, which has a consequential impact on the inferential outcome. Hence, the values of p and k should be chosen carefully depending on the contexts of investigation. For example, if the researcher strongly believes in the economic theory that the time series under investigation is stationary, one may choose a value of p close to 0. In the event that a Type II error leads to a huge loss relative to that of a Type I error, one may choose a large value of k. However, when such contexts do not dictate (as in many academic research and practical applications), the most sensible values of p and k are 0.5 and 1. The use of a conventional level of significance as a routine benchmark implies that the researcher is employing arbitrary and unknown values of p and k.

5. Concluding Remarks

This paper re-evaluates the key past results of unit root testing at the optimal level of significance chosen based on the line of enlightened judgement (Leamer, 1978). Prior studies exclusively adopt the conventional level, which by any criterion are arbitrary and not optimal. More importantly, the use of conventional level of significance completely ignores the low power associated with the unit root test. In this paper, we choose the optimal level by minimizing the expected loss from Type I and II errors, in explicit consideration of the power of the test, following Leamer (1978). When the optimal level is chosen under a symmetric loss function with an assumption that the null and alternative hypotheses are equally likely to be true, we find that the optimal levels for the ADF and DF-GLS tests are in the range of 0.2 to 0.4 for the sample sizes frequently encountered in practice. These values are well above the conventional levels, and consistent with Winer's (1962) conjecture. We also propose calibration rules based on asymptotic local power for several unit root tests, which are simple to use in practice with the value of local-to-unity coefficient as a key input.

At the optimal levels of significance, we find many time series in the extended Nelson-Plosser data set to be trend-stationary, including the real (per capita) GNP, unemployment rate, employment, and money stock. On the other hand, the price time series (consumer prices, wages, common stock prices, and GNP deflator) and nominal GDP are found to have a unit root. These findings are largely consistent with the Bayesian results of Schotman and van Dijk (1991), who find that the real variables are trend-stationary while nominal ones are difference-stationary. For the real exchange rates studied by Elliot and Pesavento (2006), both ADF and DF-GLS tests demonstrate strong support for purchasing power parity at the optimal levels of significance, in contrast with the results at a conventional level. In addition, most of the real interest rates studied by Rapach and Weber (2004) are found not to have a unit root, based on the ADF and DF-GLS tests at the optimal levels of significance.

The results obtained in this study strongly suggest that the conventional levels of significance represent a rather poor benchmark for popular unit root tests. They may be justifiable only when the researcher has a strong prior belief that the unit root is present or when the loss of Type I error disproportionately outweighs that of Type II error. We propose that empirical researchers choose the level carefully in consideration of a range of factors including the power of the test, for more sound empirical analysis, especially in the context of unit root testing. Mindless and mechanical use of the conventional levels should be avoided, as Engsted (2009) and Kim and Ji (2015) point out. In response to the univariate unit root test's low power, a number of panel unit root tests with substantially higher power have been proposed (see, for an up-to-date review, Choi, 2015). For these tests, it is highly likely that the optimal level of significance should be set at a much lower level than the conventional one. With a low probability of Type II error, a panel test at a conventional level (e.g., 0.05) may be severely biased towards the Type I error, with a consequence that a true null hypothesis is rejected too often. By lowering the level of significance, a reasonable balance between the two error probabilities can be attained (see, for a related discussion, Kim and Ji, 2015). There are also a number of unit root tests which incorporate the effects of structural breaks (see, for a review, Choi, 2015). The empirical results based on these tests may also be re-evaluated at the optimal levels of significance. We leave these lines of research as possible future research topics.

References

Campbell, J. Y. Mankiw, N. G. 1987. Are Output Fluctuations Transitory? *Quarterly Journal of Economics*, 102, 857-80.

Cheung, Y-W, Lai. K.S. 1995. Lag order and critical values of a modified Dickey-Fuller test, *Oxford Bulletin of Economics and Statistics*, 57 (3), 411-419.

Choi, I. 2015. *Almost All About Unit Roots*, Cambridge University Press, New York.

Cochrane, J. H. 1991. A critique of application of unit root tests, *Journal of Economic Dynamics and Control*, 15, 275-284.

Darné, O. 2009. The uncertain unit root in real GNP: A re-examination, *Journal of Macroeconomics*, 31(1), 153- 166.

Davidson R. MacKinnon, J.S. 1993. *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

DeJong, D.N., Nankervis, J.C., Savin, N.E., Whiteman, C.H. 1992. Integration versus trend stationary in time series, *Econometrica*, 60, 423–433.

Dickey, D.A., Fuller, W.A. 1979. Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association*, 74(366), 427–431.

Diebold, F.X., Senhadji, A. 1996. The uncertain root in real GNP: Comment, *American Economic Review*, 86, 1291–1298.

Elliott, G., Pesavento. E. 2006. On the Failure of Purchasing Power Parity for Bilateral Exchange Rates after 1973, *Journal of Money, Credit, and Banking*, 38(6), 1405-1429.

Elliott, G., Rothenberg, T.J., Stock, J.H.,1996. Efficient tests for an autoregressive unit root, *Econometrica*, 64(4), 813–836.

Engsted, T. 2009. Statistical vs. economic significance in economics and econometrics: Further comments on McCloskey and Ziliak, *Journal of Economic Methodology*, 16(4), 393-408.

Hausman, J.A. 1978. Specification Tests in Econometrics, *Econometrica*, 46(6), 1251-1271.

Keuzenkamp, H.A. Magnus, J. 1995. On tests and significance in econometrics, *Journal of Econometrics*, 67(1), 103–128.

Kilian, L. 1998. Small sample confidence intervals for impulse response functions, *The Review of Economics and Statistics*, 80, 218– 230.

Kim, J.H. 2004. Bootstrap Prediction Intervals for Autoregression using Asymptotically Mean-Unbiased Parameter Estimators, *International Journal of Forecasting*, 20, 85-97.

Kim, J. H. 2015. BootPR: Bootstrap Prediction Intervals and Bias-Corrected Forecasting. R package version 0.60. Available at http://CRAN.R-project.org/package=BootPR

Kim, J. H., Ji, P. 2015. Significance Testing in Empirical Finance: A Critical Review and Assessment, *Journal of Empirical Finance*, 34, 1-14.

Kish, L. 1959. Some statistical problems in research design, *American Sociological Review*, 24, 328-338.

Koop, G., Steel, M.F.J. 1994. A Decision-Theoretic Analysis of the Unit-Root Hypothesis Using Mixtures of Elliptical Models, *Journal of Business and Economic Statistics*, 12(1), 95-107.

Leamer, E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley, New York.

Lehmann E.L., Romano, J.S. 2005. *Testing Statistical Hypothesis*, 3rd edition, Springer, New York.

Lothian J.R., Taylor, M.P. 1996. Real exchange rate behavior: the recent float from the perspective of the past two centuries, *Journal of Political Economy*, 104, 488–510.

Luo, S., Startz, R. 2014. Is it one break or ongoing permanent shocks that explains U.S. real GDP? *Journal of Monetary Economics*, 66, 155-163.

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

MacKinnon, J. G. 2002. Bootstrap inference in Econometrics, *Canadian Journal of Economics*, 35(4), 615-644.

Maddala, G.S., Kim, I.-M. 1998. *Unit Roots, Cointegration and Structural Changes*, Cambridge University Press, Cambridge.

McCallum, B. T. 1986. On "Real" and "Sticky-Price" Theories of the Business Cycle, *Journal of Money, Credit and Banking*, 18(4), 397-414.

Morrison, D. E., Henkel, R. E. (eds.) 1970. *The Significance Test Controversy: A Reader*, Aldine Transactions, New Brunswick, NJ.

Müller, K. U., Elliott, G. 2003. Testing for unit roots and the initial condition, *Econometrica*,71(4), 1269-1286.

Murray, C.J., Nelson, C. R. 2000. The uncertain trend in U.S. GDP, *Journal of Monetary Economics*, 46, 79-95.

Nelson, C.R., Plosser, C.I., 1982. Trends and random walks in macroeconomic time series, *Journal of Monetary Economics*, 10, 139–162.

Orcutt, G. H., Winokur, H. S. 1969. First order autoregression: inference, estimation and prediction, *Econometrica*, 37, 1-14.

Papell D. 1997. Searching for stationarity: purchasing power parity under the current float, *Journal of International Economics*, 43, 313–332.

Papell, D.H., Prodan, R. 2004. The uncertain unit root in US real GDP: Evidence with restricted and unrestricted structural change, *Journal of Money, Credit and Banking*, 36, 423–427.

Pfaff, B. 2008. *Analysis of Integrated and Cointegrated Time Series with R*, 2nd edition, Springer, New York.

Phillips, P., Perron, P. 1988. Testing for a Unit Root in Time Series Regression, *Biometrika*, 75, 335-346.

Rapach, D. E., Weber, C. E. 2004. Are real interest rates really nonstationary? New evidence from tests with good size and power, *Journal of Macroeconomics*, 26(3), 409–430.

R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.

Rose, A.K. 1988. Is the real interest rate stable? *Journal of Finance,* 43(5), 1095−1112.

Rudebusch, G. 1993. The uncertain unit root in real GNP, *American Economic Review*, 83, 264–272.

Schotman, P.C. van Dijk, H.K. 1991. On Bayesian Roots to Unit Roots, *Journal of Applied Econometrics*, 6, 387-401.

Schwert, G., 1989. Testing for unit roots: A Monte Carlo investigation, *Journal of Business and Economic Statistics*, 7(2), 147−159.

Skipper, J. K., Guenther, A. L. and Nass, G. 1967. The sacredness of .05: a note on concerning the use of statistical levels of significance in social science, *The American Sociologist*, 2, 16-18.

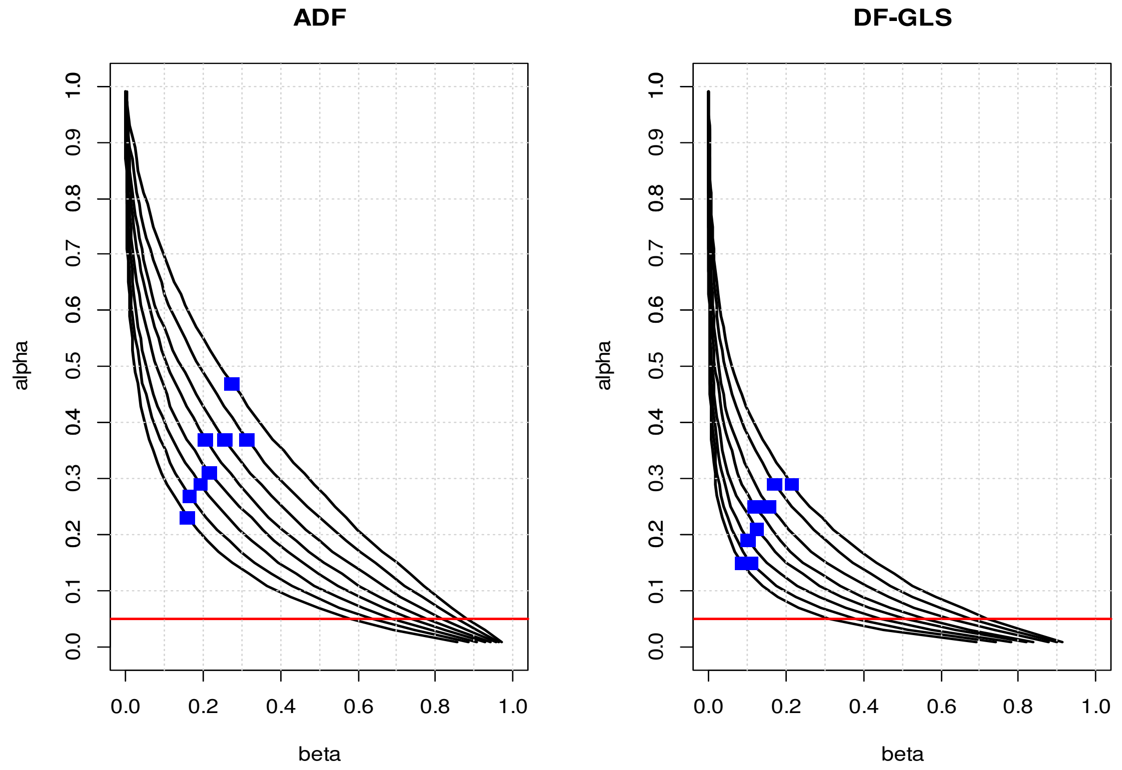
Startz, R. 2014. Choosing the More Likely Hypothesis, *Foundations and Trends in Econometrics*, 7(2), 119–189.

Stine, R. A., Shaman, P. 1989. A fixed point characterization for bias of autoregressive estimators, *The Annals of Statistics*, 17, 1275-1284.

Winer, B. J. 1962. *Statistical Principles in Experimental Design*, New York, McGraw-Hill.

Ziliak, S. T. McCloskey, D.N. 2008, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, The University of Michigan Press.

Figure 1. Lines of Enlightened Judgement (λ_1= -0.15; model with constant and linear trend)



The lines of enlightened judgement are plotted in black, corresponding to the sample sizes n = (60, 70, 80, 90, 100, 110, 120, 130) from far right to the left. The red horizontal lines correspond to α = 0.05. The square dots indicate the points where $\alpha+\beta$ is minimized.

ADF test

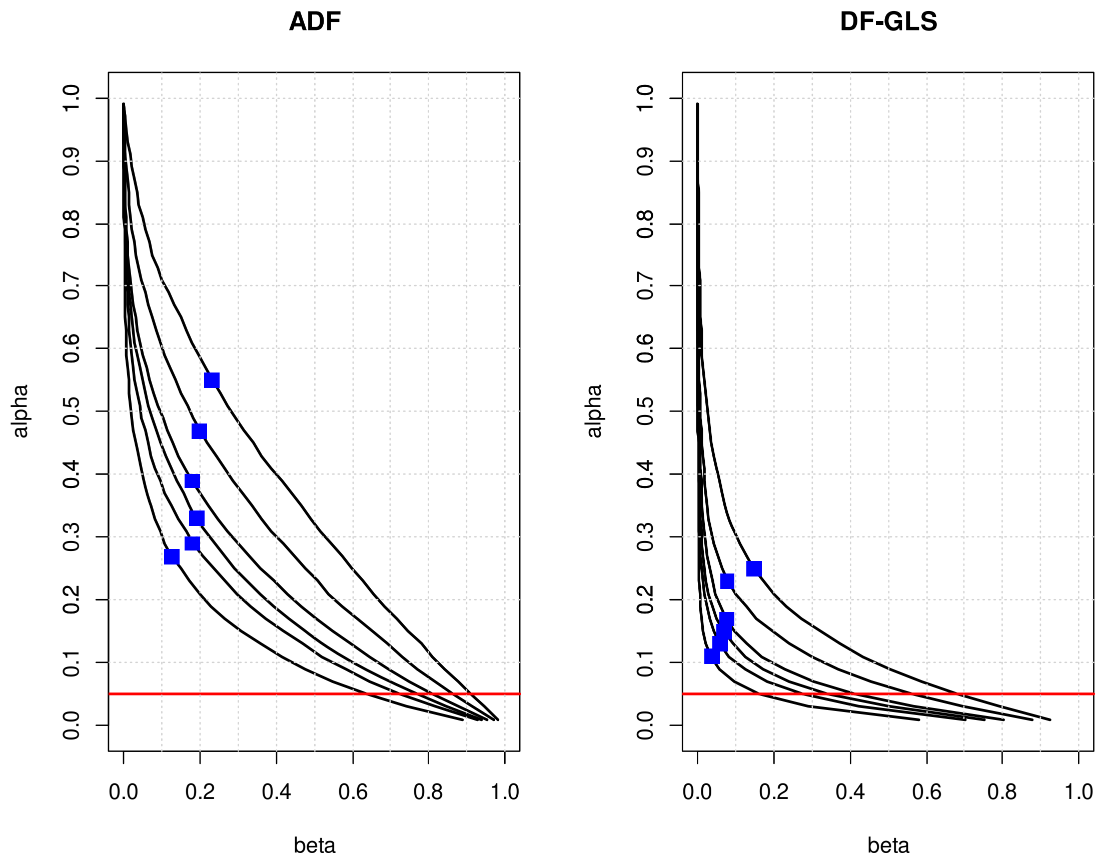
n	α	β	power	CR	α*	β*	power	CR*
	α fixed at 0.05				minimize $\alpha+\beta$			
60	0.05	0.88	0.12	-3.49	0.47	0.27	0.73	-2.22
70	0.05	0.86	0.14	-3.48	0.37	0.31	0.69	-2.41
80	0.05	0.82	0.18	-3.47	0.37	0.25	0.75	-2.41
90	0.05	0.78	0.22	-3.46	0.37	0.21	0.79	-2.41
100	0.05	0.74	0.26	-3.46	0.31	0.22	0.78	-2.54
110	0.05	0.69	0.31	-3.46	0.29	0.19	0.81	-2.58
120	0.05	0.64	0.36	-3.45	0.27	0.16	0.84	-2.63
130	0.05	0.58	0.42	-3.44	0.23	0.16	0.84	-2.72

DF-GLS test

n	α	β	power	CR	α*	β*	power	CR*
	α fixed at 0.05				minimize $\alpha+\beta$			
60	0.05	0.72	0.28	-2.89	0.29	0.21	0.79	-2.05
70	0.05	0.68	0.32	-2.89	0.29	0.17	0.83	-2.05
80	0.05	0.62	0.38	-2.89	0.25	0.15	0.85	-2.13
90	0.05	0.56	0.44	-2.89	0.25	0.12	0.88	-2.13
100	0.05	0.51	0.49	-2.89	0.21	0.12	0.88	-2.24
110	0.05	0.44	0.56	-2.89	0.19	0.10	0.90	-2.29
120	0.05	0.38	0.62	-2.89	0.15	0.11	0.89	-2.40
130	0.05	0.31	0.69	-2.89	0.15	0.08	0.92	-2.40

CR: Critical value at 5% level; CR*: Critical value associated with α*.

Figure 2. Lines of Enlightened Judgement (λ_1= -0.05; model with a constant)



The lines of enlightened judgement are plotted in black, corresponding to the sample sizes n = (80, 120, 160, 180, 200, 240) from far right to the left. The red horizontal lines correspond to α = 0.05. The square dots indicate the points where $\alpha+\beta$ is minimized.

ADF test

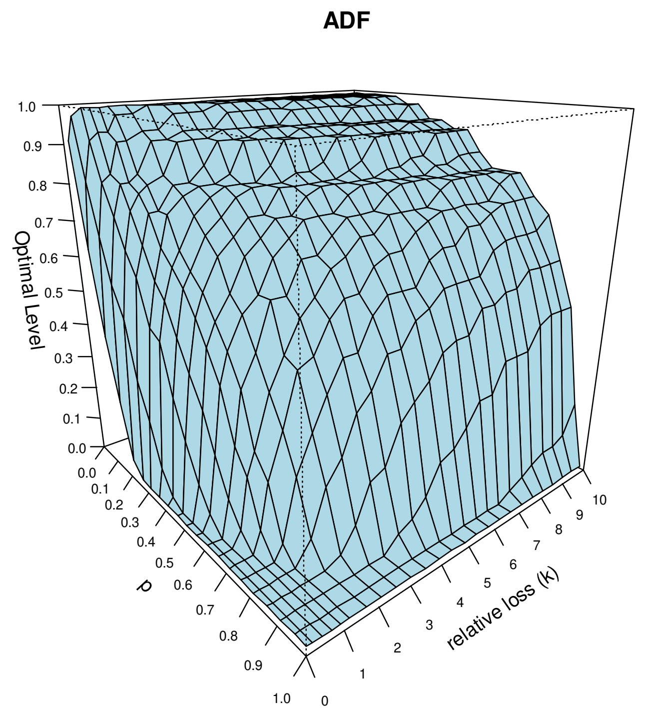
n	α	β	Power	CR	α^*	β^*	power	CR*
	α fixed at 0.05				minimize $\alpha+\beta$			
80	0.05	0.91	0.09	-2.90	0.55	0.23	0.77	-1.46
120	0.05	0.87	0.13	-2.89	0.47	0.20	0.80	-1.62
160	0.05	0.81	0.19	-2.88	0.39	0.18	0.82	-1.78
180	0.05	0.78	0.22	-2.88	0.33	0.19	0.81	-1.90
200	0.05	0.73	0.27	-2.88	0.29	0.18	0.82	-1.99
240	0.05	0.63	0.37	-2.87	0.27	0.13	0.87	-2.04

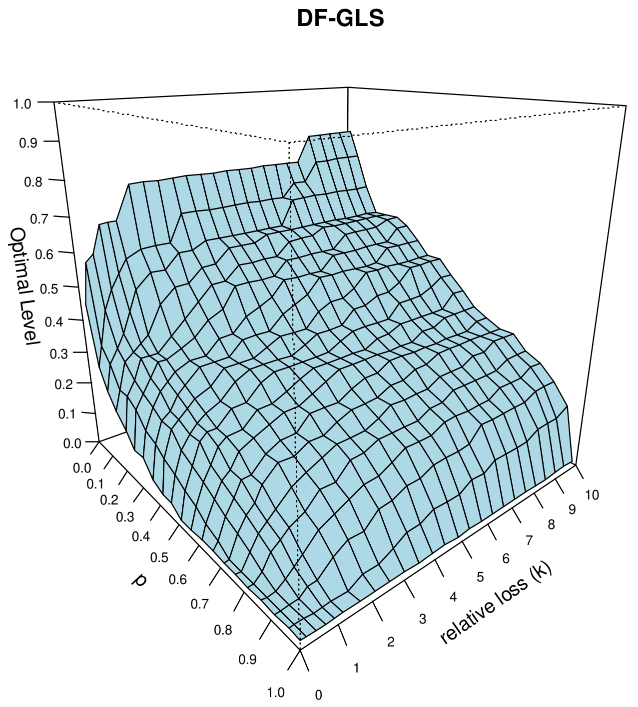
DF-GLS test

n	α	β	power	CR	α^*	β^*	power	CR*
	α fixed at 0.05				minimize $\alpha+\beta$			
80	0.05	0.68	0.32	-1.94	0.25	0.15	0.85	-1.08
120	0.05	0.56	0.44	-1.94	0.23	0.08	0.92	-1.14
160	0.05	0.41	0.59	-1.94	0.17	0.08	0.92	-1.33
180	0.05	0.34	0.66	-1.94	0.15	0.07	0.93	-1.40
200	0.05	0.27	0.73	-1.94	0.13	0.06	0.94	-1.48
240	0.05	0.16	0.84	-1.94	0.11	0.04	0.96	-1.57

CR: Critical value at 5% level; CR*: Critical value associated with α*.

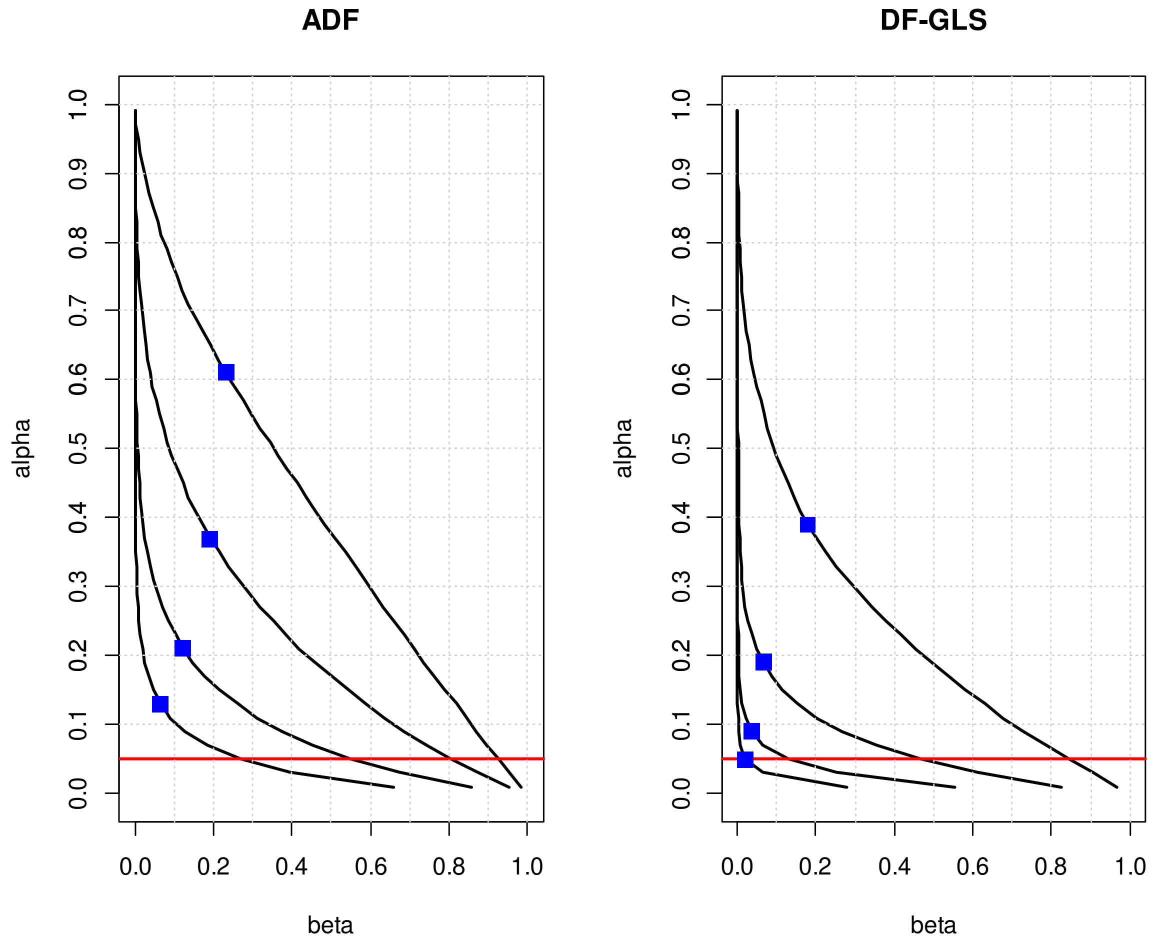
Figure 3. Optimal Level of Significance, prior probability, and relative loss





Each figure plots the optimal level of significance which minimizes the expected loss against $p = P(H_0)$ and $k = L_2/L_1$ (relative loss), for the ADF and DF-GLS tests (models with constant only) when $n = 120$. A grid of p values between 0 and 1 is used along with a grid of k values between 0 and 10. All other settings for calculation are the same as those in Figure 2.

Figure 4. Lines of Enlightened Judgement based on Asymptotic Local Power (model with a constant)



The lines of enlightened judgement are plotted in black, corresponding to the local-to-unity coefficients c = (2.5, 7.5, 12.5, 17.5) from far right to the left (n = 500). The red horizontal lines correspond to α = 0.05. The square dots indicate the points where $\alpha+\beta$ is minimized.

Table 1. Optimal level of significance, autocorrelation coefficient, and initial value

	ADF				DF-GLS			
ρ_1	-0.5	0	0.5	0.9	-0.5	0	0.5	0.9
X_0^*								
0	0.45	0.47	0.51	0.51	0.19	0.21	0.21	0.27
1	0.47	0.45	0.49	0.47	0.23	0.21	0.21	0.27
2	0.45	0.49	0.51	0.53	0.29	0.25	0.21	0.27
3	0.45	0.47	0.51	0.51	0.35	0.29	0.23	0.27
4	0.45	0.49	0.49	0.47	0.45	0.35	0.25	0.27
5	0.43	0.47	0.47	0.49	0.55	0.41	0.27	0.29
6	0.37	0.43	0.47	0.49	0.63	0.49	0.31	0.29
8	0.33	0.37	0.45	0.51	0.75	0.59	0.35	0.27
10	0.27	0.35	0.43	0.47	0.81	0.67	0.43	0.27

The entries are the optimal level of significance when $P(H_0) = 0.5$, $L_1 = L_2$, $n = 120$.

ρ_1: the coefficient of the augmentation term given in (2); X_0*: standardized starting value of (1)

Table 2. Extended Nelson-Plosser data: annual U.S. macroeconomic time series to 1988

	n	ADF			DF-GLS		
		p-value	α*	Decision*	p-value	α*	Decision*
Real GNP	80	0.05	0.37	Reject	0.05	0.25	Reject
Nominal GNP	80	0.58	0.37	Accept	0.49	0.25	Accept
Real per capital GNP	80	0.04	0.37	Reject	0.06	0.25	Reject
Industrial Production	129	0.26	0.23	Accept	0.27	0.15	Accept
Employment	99	0.18	0.31	Reject	0.04	0.21	Reject
Unemployment Rate	99	0.01	0.31	Reject	0.01	0.21	Reject
GNP deflator	100	0.70	0.31	Accept	0.73	0.21	Accept
Consumer Prices	129	0.91	0.22	Accept	0.61	0.15	Accept
Wages	89	0.53	0.37	Accept	0.37	0.25	Accept
Real Wages	89	0.75	0.37	Accept	0.51	0.25	Accept
Money Stock	100	0.18	0.31	Reject	0.10	0.21	Reject
Velocity	120	0.78	0.27	Accept	0.87	0.15	Accept
Interest Rate	89	0.98	0.37	Accept	0.33	0.25	Accept
Common Stock Prices	118	0.64	0.27	Accept	0.63	0.15	Accept

A model with constant and linear trend is used for all series; the p-values of the ADF test are obtained from the statistics reported in Schotman and van Dijk (1992) using MacKinnon's (1994) method, while that of the DF-GLS is derived from the simulated asymptotic distribution; α*: the optimal levels of significance from Figure 1. Decision*: Decision for H_0 based on α* given in Figure 1. The lag orders used are the same as those used by Nelson and Plosser (1982).

Table 3. Elliott-Pesavento data: quarterly real exchange rates from 1973 to 2003 (n=120)

	ADF			DF-GLS		
	statistic	p-value	Decision* (α*=0.47)	statistic	p-value	Decision* (α*=0.23)
Austria	-1.729	0.414	Reject	-1.155	0.225	Reject
Belgium	-2.319	0.168	Reject	-2.133*	0.032	Reject
Canada	-1.297	0.629	Accept	-0.487	0.503	Accept
Denmark	-2.507	0.116	Reject	-2.318*	0.020	Reject
Finland	-2.377	0.150	Reject	-1.847	0.061	Reject
France	-1.955	0.306	Reject	-1.965*	0.047	Reject
Germany	-1.996	0.288	Reject	-2.006*	0.043	Reject
Italy	-1.966	0.301	Reject	-1.975*	0.047	Reject
Japan	-2.265	0.185	Reject	-1.208	0.207	Reject
Netherlands	-1.755	0.401	Reject	-1.714	0.082	Reject
Norway	-2.178	0.215	Reject	-2.135*	0.032	Reject
Spain	-1.928	0.319	Reject	-1.478	0.130	Reject
Sweden	-2.219	0.201	Reject	-1.997*	0.044	Reject
Switzerland	-2.499	0.118	Reject	-1.521	0.120	Reject
UK	-2.363	0.154	Reject	-1.703	0.084	Reject

The ADF and DF-GLS statistics (model with constant only) are re-produced from Elliot and Pesavento (2006). The p-values are calculated using MacKinnon's (1994) method. The asterisk indicates the rejection of the null hypothesis of a unit root at 5% level of significance. Decision*: Decision for H_0 based on α* given in Figure 2. The critical value for the ADF test corresponding to α* = 0.47 is -1.62, while that of the DF-GLS test corresponding to α* = 0.23 is -1.14 for n = 120.

Table 4. Rapach-Weber data: quarterly real interest rates from 1957 to 2000 (n=173)

	ADF			DF-GLS		
	statistic	p-value	Decision* (α*=0.33)	statistic	p-value	Decision* (α*=0.15)
Belgium	-2.22	0.200	Reject	-1.99*	0.045	Reject
Canada	-2.12	0.237	Reject	-1.78	0.071	Reject
Denmark	-2.93*	0.044	Reject	-1.33	0.169	Accept
France	-2.08	0.253	Reject	-2.23*	0.025	Reject
Ireland	-2.35	0.158	Reject	-1.10	0.245	Accept
Italy	-2.42	0.138	Reject	-1.44	0.139	Reject
Japan	-2.49	0.120	Reject	-2.06*	0.038	Reject
Netherlands	-1.44	0.562	Accept	-0.99	0.287	Accept
New Zealand	-1.35	0.606	Accept	-1.11	0.241	Accept
UK	-2.98*	0.039	Reject	-2.64*	0.009	Reject

The ADF and DF-GLS statistics (model with constant only) are re-produced from Rapach and Weber (2004). The p-values are calculated using MacKinnon's (1994) method. The asterisk indicates the rejection of the null hypothesis of a unit root at 5% level of significance. Decision*: Decision for H_0 based on α* given in Figure 2. The critical value for the ADF test corresponding to α* = 0.33 is -1.90, while that of the DF-GLS test corresponding to α* = 0.15 is -1.40 when n = 180.