

UNIT ROOT TESTS USING MORE MOMENT CONDITIONS THAN LEAST SQUARES*

Kyung So Im
University of Central Florida

June 2001

Abstract

This paper derives the asymptotic distribution of the unit root test statistic when more moment conditions than the least squares moment conditions are used, and shows that the same asymptotic efficiency is reached through computationally simple residual augmented least squares (RALS) estimator proposed by Im and Schmidt (1999). Following Monte Carlo simulation, the small sample size of the RALS-based unit root tests is quite close to the asymptotic size; the power improves over the standard Dickey-Fuller test when the error is not normal, and is compared favorably to the other unit root tests that are designed to be more powerful than the Dickey-Fuller test for non-normal errors.

JEL Classification: C22, C12, C13.

Key Words: Unit root test, Generalized methods of moments, Residual augmented least squares.

*I am grateful to Frank Barton School of Business for its financial support.

1. Introduction

The most widely used test of the unit root hypothesis is the augmented Dickey-Fuller (ADF) test. ADF is based on least squares, and therefore has desirable properties when the time series is driven by normal variables. However, there would be more powerful tests than ADF when the innovations do not follow normal distribution. Several authors have investigated this possibility: Cox and Llatas (1991) studied the asymptotic distribution of the maximum likelihood estimators (MLE) in Dickey-Fuller regression assuming the true error density is known. Lucas (1995) derived the asymptotic distribution of the unit root statistics based on M-estimate. Herce (1996) studied the least absolute deviation-based unit root tests. Hasan and Koenker (1997) proposed rank tests. Shin and So (1999) and Beelders (1996) studied the unit root tests based on the adaptive estimation. According to the simulation results reported by these authors, the power of unit root test could be substantially improved over the Dickey-Fuller test when the data series are driven by non-normal errors.

In this paper we follow the framework of generalized methods of moments (GMM) to investigate possibly more powerful tests obtained from using moment conditions beyond those used in least squares. The asymptotic distribution of the GMM-based unit root test is derived directly from the results obtained by Lucas (1995) and Hansen(1995).

We also show that the estimator obtained from simple two step least squares procedure, proposed and referred to as residual augmented least squares (RALS) by Im and Schmidt (1999), is asymptotically identical to the asymptotic distribution of GMM under unit root environment. This is an extension of Im and Schmidt (1999), where they showed that RALS is asymptotically identical to GMM under iid environment.

According to our simulation results, the small sample size of the RALS-based unit root tests is quite close to the asymptotic size. The power generally improves over ADF when the errors are not normal, and is compared favorably to the other tests that are designed to be more powerful than ADF for non-normal errors such as the tests based on adaptive estimate or M-estimate.

2. Asymptotic Distribution of GMM Unit Root Statistics

In this section we derive the asymptotic distributions of GMM estimators resulting from least squares moments and some additional moment restrictions are used. We also derive the asymptotic distributions of their associated t-statistics.

Consider a time series that follows:

$$y_t = \phi y_{t-1} + \varepsilon_t, \quad t = 1, 2, \dots \quad (2.1)$$

where $\{\varepsilon_t\}_{t=1}^{\infty}$ is a sequence of innovations. We are interested in the test of unit

root hypothesis $H_0 : \phi = 1$ against the alternative hypothesis $H_A : \phi < 1$. We assume:

Assumption 1. $\varepsilon_t = \sum_{j=1}^p a_j \varepsilon_{t-j} + e_t$, $t = 1, 2, \dots$, where $\{e_t\}_{t=1}^\infty$ is an iid sequence with zero mean and finite second moment σ_e^2 , and all the roots of $a(z) = 1 - \sum_{j=1}^p a_j z^j$ lie outside of unit circle.

If Assumption 1 is met and $y_0 = 0$, an appropriate model is ADF:

$$\Delta y_t = \beta y_{t-1} + \sum_{j=1}^p \delta_j \Delta y_{t-j} + e_t, \quad t = 1, 2, \dots, \quad (2.2)$$

where $\Delta y_t = y_t - y_{t-1}$ and $\beta = \phi - 1$. Let $\hat{\beta}_{LS}$ be the least squares estimator of β in regression (2.2), and t_{LS} be its associate t-statistic. Then, it is well known, under the null hypothesis,

$$T \hat{\beta}_{LS} \Rightarrow a(1) \left(\int_0^1 [W(r)]^2 dr \right)^{-1} \int_0^1 W(r) dW(r), \quad (2.3)$$

and

$$t_{LS} \Rightarrow \left(\int_0^1 [W(r)]^2 dr \right)^{-1/2} \int_0^1 W(r) dW(r) \equiv DF, \quad (2.4)$$

where $a(1) = 1 - \sum_{j=1}^p a_j$, and $W(r)$ denotes the standard Brownian motion on $r \in [0, 1]$.

Let $\xi_t = (\Delta y_{t-1}, \Delta y_{t-2}, \dots, \Delta y_{t-p})'$, and $z_t = (y_{t-1}, \xi_t)'$. Suppose we have $J \times (p+1)$ additional moment conditions:

$$E [g(e_t) \otimes z_t] = 0, \quad t = 1, 2, \dots, \quad (2.5)$$

where $g(e_t)$ is $J \times 1$ vector that satisfies:

Assumption 2. $g(\cdot)$ is differentiable and satisfies the first-order Lipschitz condition: $|g'_j(x) - g'_j(y)| < M |x - y|$ for some constant M for all j , where $g_j(\cdot)$ is the j -th element of $g(\cdot)$. $E [g(e_t)] = 0$, the second moment of $g(e_t)$ exists, and $E [g'(e_t)] < \infty$.

Define $C = E [g(e_t)g(e_t)']$ and $D = E [g'(e_t)]$, and $\psi(e_t) = D'C^{-1}g(e_t)$, for $t = 1, 2, \dots$. Also define the correlation between e_t and $\psi(e_t)$ by

$$\rho = \frac{\sigma_{\psi e}}{\sigma_\psi \sigma_e} \quad (2.6)$$

where $\sigma_\psi^2 = Var [\psi(e_t)] = Var [D'C^{-1}g(e_t)] = D'C^{-1}D$, $\sigma_{\psi e} = E [\psi(e_t)e_t] = DC^{-1}E [g(e_t)e_t]$.

Theorem 1. Suppose a time series follows (2.1), and Assumptions 1 and 2 are satisfied. Under the null hypothesis, we have, for, $\tilde{\beta}_G$, GMM estimator using the moments conditions (2.5) in the ADF regression (2.2);

$$T\tilde{\beta}_G \Rightarrow \frac{a(1)}{\sigma_e\sigma_\psi} \left(\int_0^1 [W_1(r)]^2 dr \right)^{-1} \int_0^1 W_1(r)dW_2. \quad (2.7)$$

Also, for its t-statistic obtained by $t_G = \tilde{\beta}_G/se(\tilde{\beta}_G)$, where

$$se(\tilde{\beta}_G) = \tilde{\sigma}_\psi^{-1} \sqrt{\left(\sum_{t=1}^T y_{t-1}^2 - \sum_{t=1}^T y_{t-1}\xi_t \left(\sum_{t=1}^T \xi_t\xi_t' \right)^{-1} \sum_{t=1}^T \xi_t'y_{t-1} \right)^{-1}},$$

$\tilde{\sigma}_\psi^2 = \tilde{D}'\tilde{C}^{-1}\tilde{D}$, $\tilde{D} = T^{-1}\sum_{t=1}^T g'(\tilde{e}_t)$, $\tilde{C} = T^{-1}\sum_{t=1}^T g(\tilde{e}_t)g(\tilde{e}_t)'$, and \tilde{e}_t is the residual from GMM estimation in the regression (2.2), we then have

$$t_G \Rightarrow \rho DF + \sqrt{1 - \rho^2}N(0, 1), \quad (2.8)$$

where ρ is defined in (2.6), DF denotes the Dickey-Fuller distribution as was defined in (2.4), and $N(0, 1)$ signifies the standard normal distribution.

proof. See Appendix.

In case when an intercept is allowed in the regression (2.2);

$$\Delta y_t = \alpha_1 + \beta y_{t-1} + \sum_{j=1}^p \delta_j \Delta y_{t-j} + e_t, \quad t = 1, 2, \dots, \quad (2.9)$$

we have the additional moment conditions $E[g(e_t) \otimes (1, z_t)'] = 0$. In view of the expression for β estimator in (A.9) of Appendix, we will have the GMM estimator

$$T\tilde{\beta}_{G,\mu} = \left(\sigma_\psi^2 T^{-2} \sum_{t=1}^T \tilde{y}_{t-1}^2 \right)^{-1} T^{-1} \sum_{t=1}^T \tilde{y}_{t-1} \psi(e_t) + o_p(1),$$

where $\tilde{y}_{t-1} = y_{t-1} - T^{-1}\sum_{t=1}^T y_{t-1}$, $t = 1, 2, \dots$. Consequently,

$$T\tilde{\beta}_{G,\mu} \Rightarrow \frac{a(1)}{\sigma_\psi\sigma_e} \int_0^1 \tilde{W}_1(r)dW_2(r) / \int_0^1 [\tilde{W}_1(r)]^2 dr, \quad (2.10)$$

where $\tilde{W}_1(r)$ is the demeaned Brownian motion; $\tilde{W}_1(r) = W_1(r) - \int_0^1 W_1(r)dr$. Also by construction

$$t_{G,\mu} \Rightarrow \rho DF_\mu + \sqrt{1 - \rho^2}N(0, 1),$$

where DF_μ denotes the limiting distribution of the t-statistic of the least squares in regression (2.9).

Similarly, when the model includes a linear time trend as well as an intercept we have the model:

$$\Delta y_t = \alpha_1 + \alpha_2 t + \beta y_{t-1} + \sum_{j=1}^p \delta_j \Delta y_{t-j} + e_t, \quad t = 1, 2, \dots, \quad (2.11)$$

and will have for GMM estimator:

$$T\tilde{\beta}_{G,\tau} \Rightarrow \frac{a(1)}{\sigma_\psi \sigma_e} \int_0^1 \check{W}_1(r) d\check{W}_2(r) / \int_0^1 [\check{W}_1(r)]^2 dr, \quad (2.12)$$

where $\check{W}(r)$ is the detrended Brownian motion. Also,

$$t_{G,\tau} \Rightarrow \rho DF_\tau + \sqrt{1 - \rho^2} N(0, 1), \quad (2.13)$$

where DF_τ denotes the limiting distribution of the t-statistic for the OLS estimator of β in regression (2.11).

Remark 1. Asymptotic distribution of t_G depends on the nuisance parameter ρ . Hansen (1995) reports the critical values of the asymptotic distribution of the t-statistics for $\rho^2 = 0.1$ to 1.0 at step of 0.1.

3. RALS Estimation

Consider the model (2.9), the ADF model with an intercept. Suppose $g(e_t) = (e_t, [h(e_t) - K]')'$. Let $x_t = (1, z_t)'$. Then, we have the moment conditions; $E[g(e_t) \otimes x_t] = 0$, which we split into the least squares moment conditions;

$$E(e_t \otimes x_t) = 0 \quad (3.1)$$

and additional $2(J - 1)$ moment conditions;

$$E[(h(e_t) - K) \otimes x_t] = 0. \quad (3.2)$$

Therefore, we have

$$C = \begin{bmatrix} \sigma_e^2 & C'_{21} \\ C_{21} & C_{22} \end{bmatrix}, \quad \text{and} \quad D = \begin{bmatrix} 1 \\ D_2 \end{bmatrix}, \quad (3.3)$$

where $C_{21} = E[e_t h(e_t)]$, $C_{22} = E[h(e_t)h(e_t)']$, and $D_2 = E[h'(e_t)]$.

Let

$$\hat{w}_t = h(\hat{e}_t) - \hat{K} - \hat{e}_t \hat{D}_2, \quad t = 1, 2, \dots, \quad (3.4)$$

where \hat{e}_t is the OLS residual from the regression (2.9), $\hat{K} = \frac{1}{T} \sum_{t=1}^T h(\hat{e}_t)$, $\hat{D}_2 = \frac{1}{T} \sum_{t=1}^T h'(\hat{e}_t)$. The RALS is the least squares in the regression:

$$\Delta y_t = \alpha_1 + \beta y_{t-1} + \sum_{j=1}^p \delta_j \Delta y_{t-j} + \hat{w}_t' \gamma + \eta_t, \quad t = 1, 2, \dots \quad (3.5)$$

We show in the following that the RALS estimator is asymptotically identical to the GMM estimator using the moment conditions (3.1) and (3.2).

Theorem 2. Consider a time series described in (2.1) with $\phi = 1$. Under Assumptions 1 and 2, the RALS estimator of β , obtained by least squares in the regression (3.5), is asymptotically identical to the GMM estimator using the moment conditions (3.1) and (3.2). Also, the limiting distributions of the t-statistics are the same.

proof. See Appendix.

When there is a linear time trend included in the regression, we have

$$\Delta y_t = \alpha_1 + \alpha_2 t + \beta y_{t-1} + \sum_{j=1}^p \delta_j \Delta y_{t-j} + \hat{w}_t' \gamma + \eta_t, \quad t = 1, 2, \dots \quad (3.6)$$

By construction, we will obtain the same results as those in (2.12) and (2.13) for estimator of β and for its t-statistic.

We provide some guidance on how to apply RALS in practice:

- ρ^2 is estimated by

$$\hat{\rho}^2 = \hat{\sigma}_A^2 / \hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is the usual error variance estimator in standard ADF regression, and $\hat{\sigma}_A^2$ is the error variance estimator in RALS regression. See the proof of Theorem 2 [equations (A.16) and (A.19)]. Based on this value of $\hat{\rho}^2$, critical values are found in Hansen (1995).

- When the sample size is small (e.g. $T \leq 50$), impose the restriction of $\beta = 0$ in the first step regression that yields the residuals for augmented variables \hat{w}_t . According to our simulation experience, this procedure improves the size property of the test significantly with only minimal effects on the power. When the sample is relatively big (e.g., $T = 100$), this effect, however, disappears quickly.

4. Simulation Results

In this section we investigate the small sample property of the unit root tests based on RALS. As was noted above, we use the critical values reported by Hansen (1995) for RALS, and impose the restriction $\beta = 0$ when we construct the augmented variable. For example, suppose we have a regression model: $\Delta y_t = \alpha + \beta y_{t-1} + \sum_{j=1}^p \delta_j \Delta y_{t-j} + e_t$, $t = 1, 2, \dots$. In the first step, we estimate α and δ 's imposing $\beta = 0$, so from the regression $\widehat{\Delta y}_t = \hat{\alpha} + \sum_{j=1}^p \hat{\delta}_j \Delta y_{t-j}$, then construct the augmented variable \hat{w}_t as a function of the residuals; $\hat{e}_t = \Delta y_t - \hat{\alpha} - \sum_{j=1}^p \hat{\delta}_j \Delta y_{t-j}$. Then, in the second step, t-statistic is computed following the usual standard routine from the RALS regression: $\Delta y_t = \tilde{\alpha} + \tilde{\beta} y_{t-1} + \sum_{j=1}^p \tilde{\delta}_j \Delta y_{t-j} + \tilde{\gamma} \hat{w}_t + \hat{v}_t$.

We include two RALS estimators, RALS(2&3) and RALS(t5), in our Monte Carlo study. First, RALS(2&3) imposes additional moment conditions that the second and third moments of the errors are not correlated with the lagged dependent variables. Therefore, $h(\hat{e}_t) = [\hat{e}_t^2, \hat{e}_t^3]'$. Letting $m_j = T^{-1} \sum_{t=1}^T \hat{e}_t^j$, for $j = 2, 3$, we have for RALS(2&3);

$$\hat{w}_t = [\hat{e}_t^2 - m_2, \hat{e}_t^3 - m_3 - 3m_2 \hat{e}_t]', \quad t = 1, 2, \dots \quad (4.1)$$

The moment condition $E[(e_t^2 - \sigma_e^2) y_{t-1}] = 0$ is the condition of no heteroskedasticity, and improves the efficiency of β estimator when the errors are not symmetric, and the restriction on the third moments conditional y_{t-1} improves the efficiency unless $\mu_4 = 3\sigma^4$. See Im and Schmidt (1999) for details.

Second, RALS(t5) imposes the restrictions that arise from the score of the maximum likelihood estimation when the error density is assumed to be the t-distribution with 5 degrees of freedom. Assumption of student-t error density is a widely accepted strategy in M-estimate, and known to lead to more efficient estimation than OLS when the true density has fat-tails. Because RALS(t5) also uses the least squares restrictions, the RALS(t5) would be as efficient as least squares when the error is normally distributed, and is more efficient when the density of the errors have fat-tails. In this case, we have $h(e_t) = (c+1)e_t/(c+e_t^2)$, so $D_2 = (c+1)(c-e_t^2)/(c+e_t^2)^2$, where $c = 5$. Therefore,

$$\hat{w}_t = \frac{6\hat{e}_t}{5 + \hat{e}_t^2} - \frac{1}{T} \sum_{t=1}^T \frac{6\hat{e}_t}{5 + \hat{e}_t^2} - \hat{e}_t \frac{1}{T} \sum_{t=1}^T \frac{6(5 - \hat{e}_t^2)}{(5 + \hat{e}_t^2)^2} \quad (4.2)$$

There is no compelling reason behind of choosing $c = 5$. But, it seems that the tests are quite robust to the selection of c values. For example, following our unreported simulations, the empirical size and power of the tests are almost identical when we use $c = 3$.

We report the rejection ratio at $\alpha = 0.05$ when $\phi = 1$ to see the size property, and $\phi = 0.9$ to examine the power. We simulated the sample cases $T = 50$ and 100. All the results are based on 5,000 replications.

Table 1 reports the results for the basic case when the errors, ε_t in (2.1), are serially independent, and p , the number of ADF augmentation in the regression, is set to zero. We compare RALS(2&3) and RALS(t5) to DF, AD and M5, where DF denotes the standard Dickey-Fuller test based on OLS, AD denotes the test based on adaptive estimation studied by Shin and So (1999) and Beelders (1996), and M5 signifies the test based on M-estimate assuming the true density is student-t with 5 degrees of freedom, which was studied by Lucas (1995). The figures for AD and M5 have been reproduced from Shin and So (1999). We replicated the four distributions simulated by Shin and So (1999): (i) standard normal, (ii) t-distribution with $df = 3$, (iii) mixture normal; $0.5N(-3,1)+0.5N(3,1)$, (iv) chi-square with $df = 1$.

As is seen in Table 1, the sizes of the tests based on RALS(2&3) and RALS(t5) are quite close to the nominal 5% throughout, and the power gain over the standard DF test is substantial when the errors are not normal. The overall power of RALS(2&3) and RALS(t5) is compared favorably to the power of AD or M5. The performance of RALS(t5) and M5 are similar when the true density is the student-t with 3 degrees of freedom, but RALS(t5) is better when the density is mixture normal. When the true density is chi-square distribution with one degree of freedom, RALS(2&3), which explicitly uses the moment condition that is useful when the error is not symmetric, dominates the other tests. The AD-based test does not seem to capture the possible efficiency gain from non-symmetric feature of error density. In our simulated distributions, RALS(t5) is marginally better than RALS(2&3) when the density is symmetric. However, as we can see for the case when the density is chi-square with one degree of freedom, RALS(2&3) is generally better than RALS (t5) when the error density is skewed, and the difference often is quite substantial.

In Tables 2-5, we compare the performance of the tests when the errors are serially correlated. In doing so, we compare only of the three tests; ADF, RALS(2&3) and RALS(t5) in two data generation processes:

$$AR : \varepsilon_t = 0.5\varepsilon_{t-1} + e_t, \quad t = 1, 2, \dots,$$

and

$$MA : \varepsilon_t = e_t - 0.5e_{t-1}, \quad t = 1, 2, \dots$$

We report the size and power for fixed ADF augmentation at $p = 2$ and $p = 4$ when $T = 50$, and $p = 3$ and $p = 6$ when $T = 100$ as well as when p is selected by information criterion. We simulated Akaike and Schwarz criteria, but report only the results from Schwarz criterion. The results from Akaike criterion were similar. The minimum and maximum value of p are set 2 and 4 when $T = 50$, and 3 and 6 when $T = 100$. We consider the case when the errors are generated from standard normal, Cauchy, student-t distribution with 2 degrees of freedom, double exponential, chi-square distribution with 4 degrees of freedom, and beta(2,2) distribution. Cauchy and the t-distribution with 2

degrees of freedom do not satisfy the Assumptions 1 and 2, so that we do not know the asymptotic distributions of the statistics in this case. However, it is interesting to see the performance of the tests in this situation.

Table 2 reports the size and power of the tests for AR(1) errors when time trend is not included in ADF regressions, and Table 3 contains the case when a linear time trend is allowed. Sizes of all of the three tests reported both in Tables 2 and 3 are close to the 5% nominal size in general even when the errors are generated from Cauchy or t-distribution with two degrees of freedom. Only exception is RALS(t5). The empirical size of RALS(t5)-based test is 10-12% when the errors are from Cauchy and a time trend is allowed in the ADF regression. This result is somehow puzzling, especially because the empirical size of RALS(t5)-based test size is close to the nominal size when there is no time trend included in the regression.

The power difference between the two tests based on OLS and RALS is the greatest when the errors are generated from Cauchy. Also, as we observed in Table 1, RALS(t5) is more powerful than RALS(2&3) for all the symmetric distributions. However, RALS(2&3) is in general powerful when the errors are asymmetric. Especially, it is seen that the power of RALS(t5)-based test is lower than that of OLS-based tests when the error is chi-square 4 degrees of freedom: the power of RALS(2&3) is 52% while the power of RALS(t5) is 15% in Table 3 when a time trend is included, $T = 100$ and $p = 3$.

Tables 4 and 5 contain the results when the errors follow MA(1). When p is decided by Schwarz criterion, all the tests tend to over-reject the null hypothesis. But, when p is fixed at 4 for $T = 50$, and at $p = 6$ for $T = 100$, the size of the tests are quite close to the 5% nominal size, except when the density is Cauchy and the regression includes time trend. But, the overall size of RALS(2&3) seems as robust as the size of the standard ADF test. For the power of the tests, we observe a similar pattern as the case of AR(1) errors. RALS-based tests are substantially more powerful than OLS-based ADF tests, and RALS(2&3) is compared favorably to RALS(t5).

5. Concluding Remarks

In this paper we investigate the asymptotic distribution of GMM estimators and a simple RALS procedure of obtaining the GMM estimators. Size and power property of RALS unit root tests is investigated through simulation study. An obvious advantage of RALS over the other robust estimation procedures is in its computational simplicity. Also it turns out that the small sample size and power property of the RALS-based tests are compared favorably to the other available tests that are more powerful than ADF when the time series is driven by non-normal errors.

A. Appendix

Lemma A1. $z_t = (y_{t-1}, \xi_t)'$ as was defined before equation (2.5), and C and D are defined in (3.3). Let $(p+1) \times (p+1)$ matrix $\Upsilon_T = \text{diag}(T, \sqrt{T}, \dots, \sqrt{T})$. Then, we have, under Assumptions 1 and 2, and under the null hypothesis,

$$\sum_{t=1}^T [g'(e_t) \otimes \Upsilon_T^{-1} z_t z_t' \Upsilon_T^{-1}] \Rightarrow D \otimes \int z z', \quad (\text{A.1})$$

$$\sum_{t=1}^T g(e_t) g(e_t)' \otimes \Upsilon_T^{-1} z_t z_t' \Upsilon_T^{-1} \Rightarrow C \otimes \int z z', \quad (\text{A.2})$$

where $\int z z' = \text{diag} \left(a(1)^{-2} \sigma_e^2 \int_0^1 [W_1(r)]^2 dr, E(\xi_t \xi_t') \right)$. Also

$$\sum_{t=1}^T \psi(e_t) \Upsilon_T^{-1} z_t = \begin{bmatrix} T^{-1} \sum_{t=1}^T \psi(\varepsilon_t) y_{t-1} \\ T^{-1/2} \sum_{t=1}^T \psi(\varepsilon_t) \xi_t \end{bmatrix} \Rightarrow \begin{bmatrix} \frac{\sigma_\psi \sigma_\varepsilon}{a(1)} \int_0^1 W_1(r) dW_2(r) \\ \Gamma \end{bmatrix}, \quad (\text{A.3})$$

where $[W_1(r), W_2(r)]'$ is a bivariate Brownian motion with correlation ρ , and Γ is $p \times p$ multivariate normal variable with the covariance matrix $\sigma_\psi^2 E(\xi_t \xi_t')$.

proof. Lucas (1995, Lemma 1 in Appendix). Also see Hansen (1995, Lemma).

Lemma A2. ρ is defined in equation (2.6). Then,

$$\rho = \frac{1}{\sigma_e \sigma_\psi}. \quad (\text{A.4})$$

Also,

$$\frac{1}{\sigma_\psi^2} = \sigma_e^2 - (C_{21} - \sigma_e^2 D_2)' (C_{22} + \sigma_e^2 D_2 D_2' - C_{21} D_2' - D_2 C_{21}')^{-1} (C_{21} - \sigma_e^2 D_2). \quad (\text{A.5})$$

proof. The first result follows from a routine matrix algebra using partitioned inverse lemma. For the second result, we have, after a straightforward algebra,

$$(D' C^{-1} D)^{-1} = \sigma_e^2 \left(1 + (C_{21} - \sigma_e^2 D_2)' (\sigma_e^2 C_{22} - C_{21} C_{21}')^{-1} (C_{21} - \sigma_e^2 D_2) \right)^{-1},$$

which however is the same as $1/\sigma_\psi^2$ from Amemiya (1985, p461, Lemma 20).

PROOF OF THEOREM 1: We note the entire proof follows immediately from Lucas (1995, Theorem 1) since GMM estimator is obtained by solving the score $\sum_{t=1}^T [DC^{-1}g(e_t)z_t] = \sum_{t=1}^T [\psi(e_t)z_t] = 0$, and this score could be thought as that of M-estimation. But, we provide more details. Let $\theta = (\beta, \delta_1, \delta_2, \dots, \delta_p)'$. GMM estimator is obtained by solving the problem

$$\min_{\theta} \sum_{t=1}^T [g(e_t) \otimes z_t]' \hat{\Lambda}^{-1} \sum_{t=1}^T [g(e_t) \otimes z_t], \quad (\text{A.6})$$

where $\hat{\Lambda} = \left(\sum_{t=1}^T g(\hat{e}_t)g(\hat{e}_t)' \otimes z_t z_t' \right)$, and \hat{e}_t is residual from an initial consistent estimator of θ . Taking derivative with respect to θ , we obtain the score:

$$\sum_{t=1}^T [g'(\tilde{e}_t) \otimes z_t z_t']' \hat{\Lambda}^{-1} \sum_{t=1}^T [g(\tilde{e}_t) \otimes z_t] = 0, \quad (\text{A.7})$$

where $\tilde{e}_t = \Delta y_t - z_t \tilde{\theta}$, and $\tilde{\theta}$ is the GMM estimator. Note that $\Upsilon_T = \text{diag} \left(T, \sqrt{T}, \dots, \sqrt{T} \right)$.

Taylor expansion of the term $\sum_{t=1}^T [g(\tilde{e}_t) \otimes z_t]$ with respect to the true disturbance e_t and premultiplication of $I_J \otimes \Upsilon_T^{-1}$ yield

$$\begin{aligned} & \sum_{t=1}^T [g(\tilde{e}_t) \otimes \Upsilon_T^{-1} z_t] \\ &= \sum_{t=1}^T \left[g(e_t) \otimes \Upsilon_T^{-1} z_t - g'(e_t) \otimes \Upsilon_T^{-1} z_t z_t' \Upsilon_T^{-1} \Upsilon_T (\tilde{\theta} - \theta) \right] + o_p(1). \end{aligned} \quad (\text{A.8})$$

Solving (A.7) with respect to $\Upsilon_T (\tilde{\theta} - \theta)$, after substituting (A.8) into (A.7), we obtain:

$$\begin{aligned} \Upsilon_T (\tilde{\theta} - \theta) &= \\ & \left\{ \sum_{t=1}^T [g'(\tilde{e}_t) \otimes \Upsilon_T^{-1} z_t z_t' \Upsilon_T^{-1}]' \left[\sum_{t=1}^T g(\hat{e}_t)g(\hat{e}_t)' \otimes \Upsilon_T^{-1} z_t z_t' \Upsilon_T^{-1} \right]^{-1} \sum_{t=1}^T [g'(e_t) \otimes \Upsilon_T^{-1} z_t z_t' \Upsilon_T^{-1}] \right\}^{-1} \\ & \times \left\{ \sum_{t=1}^T [g'(\tilde{e}_t) \otimes \Upsilon_T^{-1} z_t z_t' \Upsilon_T^{-1}]' \left[\sum_{t=1}^T g(\hat{e}_t)g(\hat{e}_t)' \otimes \Upsilon_T^{-1} z_t z_t' \Upsilon_T^{-1} \right]^{-1} \sum_{t=1}^T [g(e_t) \otimes \Upsilon_T^{-1} z_t] \right\} + o_p(1). \end{aligned} \quad (\text{A.9})$$

Noting that

$$\sum_{t=1}^T \{ [g'(\tilde{e}_t) - g'(e_t)] \otimes \Upsilon_T^{-1} z_t z_t' \Upsilon_T^{-1} \} = o_p(1),$$

and

$$\sum_{t=1}^T \{ [g(\hat{e}_t)g(\hat{e}_t)' - g(e_t)g(e_t)'] \otimes \Upsilon_T^{-1} z_t z_t' \Upsilon_T^{-1} \} = o_p(1),$$

we have, from Lemma A1,

$$T\tilde{\beta}_G \Rightarrow \frac{a(1)}{\sigma_\psi\sigma_e} \left(\int_0^1 [W_1(r)]^2 dr \right)^{-1} \int_0^1 W_1(r) dW_2(r), \quad (\text{A.10})$$

where $[W_1(r), W_2(r)]$ is the bivariate Brownian motion with correlation ρ . We have for the t-statistic;

$$t_G \Rightarrow \left(\int_0^1 [W_1(r)]^2 dr \right)^{-1/2} \int_0^1 W_1(r) dW_2(r), \quad (\text{A.11})$$

which is a mixture of Dickey-Fuller and standard normal described in (2.8). To see this, note that

$$T^{-1/2} \sum_{t=1}^{[rT]} \begin{bmatrix} e_t \\ \psi(e_t) \end{bmatrix} \Rightarrow \begin{bmatrix} \sigma_e W_1(r) \\ \sigma_\psi W_2(r) \end{bmatrix}, \quad (\text{A.12})$$

where $[rT]$ denotes the integer part of rT . Therefore,

$$W_2(r) = \rho W_1(r) + \sqrt{1 - \rho^2} W_3(r), \quad (\text{A.13})$$

where $W_3(r)$ is independent of $W_1(r)$. The result follows if we note that

$$\left(\int_0^1 [W_1(r)]^2 d(r) \right)^{-1/2} \int_0^1 W_1(r) dW_3(r)$$

is standard normal.

PROOF OF THEOREM 2: Define a variable as a function of true disturbances:

$$w_t = h(e_t) - K - e_t D_2, \quad t = 1, 2, \dots$$

w_t are not observable, but we momentarily assume that they are observed. Then we show that the augmentation of w_t or \hat{w}_t yields the asymptotically same estimator of $T\beta$. Consider a regression:

$$\Delta y_t = \alpha_1 + \beta y_{t-1} + \sum_{j=1}^p \delta_j \Delta y_{t-j} + w_t' \gamma + v_t, \quad t = 1, 2, \dots \quad (\text{A.14})$$

Therefore,

$$e_t = w_t' \gamma + v_t, \quad t = 1, 2, \dots \quad (\text{A.15})$$

Let $\hat{\beta}_A^*$ be the least squares estimator of β from regression (A.14), $\sigma_v^2 = Var(v_t)$, and

$$\lambda = \frac{\sigma_{ev}}{\sigma_e \sigma_v} = \frac{\sigma_v}{\sigma_e}, \quad (\text{A.16})$$

where $\sigma_{ev} = E(\varepsilon_t v_t)$. The second equality of (A.16) follows since w_t and v_t are not correlated, so that $\sigma_{ev} = \sigma_v^2$. From Hansen (1995, Theorem 2 and 3), we have

$$T\hat{\beta}_A^* \Rightarrow \frac{\sigma_v}{\sigma_e} \left(\int_0^1 [W_4(r)]^2 \right)^{-1} \int_0^1 W_4(r) dW_5(r), \quad (\text{A.17})$$

and for the t-statistic

$$t_A^* = \lambda DF_\mu + \sqrt{1 - \lambda^2} N(0, 1), \quad (\text{A.18})$$

where $[W_4(r), W_5(r)]'$ is the bivariate Brownian motion with the correlation λ . Next, we will show that

$$\rho = \lambda. \quad (\text{A.19})$$

Note $\gamma = E(w_t w_t')^{-1} E(w_t e_t)$, so we have

$$\sigma_v^2 = \sigma_e^2 - E(e_t w_t') E(w_t w_t')^{-1} E(w_t e_t). \quad (\text{A.20})$$

Also, $E(w_t e_t) = C_{21} - \sigma_\varepsilon^2 D_2$ and $E(w_t w_t') = C_{22} + \sigma_\varepsilon^2 D_2 D_2' - C_{21} D_2' - D_2 C_{21}'$. Therefore,

$$\sigma_v^2 = \sigma_e^2 - (C_{21} - \sigma_\varepsilon^2 D_2)' (C_{22} + \sigma_\varepsilon^2 D_2 D_2' - C_{21} D_2' - D_2 C_{21}')^{-1} (C_{21} - \sigma_\varepsilon^2 D_2),$$

which is, from Lemma A1, $1/\sigma_\psi^2$. Therefore, $\rho = \lambda$.

Now we let $\hat{\beta}_A$ be the OLS estimator of β in the regression (3.5). Proof is complete if we show that $T\hat{\beta}_A$ and $T\hat{\beta}_A^*$ are identical asymptotically. Let $\hat{\zeta}_t = (\tilde{\xi}_t', \hat{w}_t')$, where $\tilde{\xi}_t = \xi_t - T^{-1} \sum_{t=1}^T \xi_t$. Then we have;

$$T\hat{\beta}_A = \frac{T^{-1} \left(\sum_{t=1}^T \tilde{y}_{t-1} e_t - \sum_{t=1}^T \tilde{y}_{t-1} \zeta_t' \left(\sum_{t=1}^T \tilde{\zeta}_t \tilde{\zeta}_t' \right)^{-1} \sum_{t=1}^T \tilde{\zeta}_t e_t \right)}{T^{-2} \left(\sum_{t=1}^T \tilde{y}_{t-1}^2 - \sum_{t=1}^T \tilde{y}_{t-1} \zeta_t' \left(\sum_{t=1}^T \tilde{\zeta}_t \tilde{\zeta}_t' \right)^{-1} \sum_{t=1}^T \zeta_t \tilde{y}_{t-1} \right)},$$

Since $T^{-1} \sum_{t=1}^T \hat{w}_t \xi_t' = o_p(1)$, and $T^{-1} \sum_{t=1}^T \tilde{\xi}_t e_t = o_p(1)$, we have:

$$T\hat{\beta}_A = \frac{T^{-1} \left(\sum_{t=1}^T \tilde{y}_{t-1} e_t - \sum_{t=1}^T \tilde{y}_{t-1} \hat{w}_t' \left(\sum_{t=1}^T \hat{w}_t \hat{w}_t' \right)^{-1} \sum_{t=1}^T \hat{w}_t' e_t \right)}{T^{-2} \left(\sum_{t=1}^T \tilde{y}_{t-1}^2 \right)} + o_p(1).$$

Similarly,

$$T\hat{\beta}_A^* = \frac{T^{-1} \left(\sum_{t=1}^T \tilde{y}_{t-1} e_t - \sum_{t=1}^T \tilde{y}_{t-1} w_t' \left(\sum_{t=1}^T \tilde{w}_t \tilde{w}_t' \right)^{-1} \sum_{t=1}^T \tilde{w}_t' e_t \right)}{T^{-2} \left(\sum_{t=1}^T \tilde{y}_{t-1}^2 \right)} + o_p(1),$$

$T\hat{\beta}_A$ and $T\hat{\beta}_A^*$ are asymptotically identical if $T^{-1} \sum \tilde{y}_{t-1} (\hat{w}_t - w_t) = o_p(1)$. However,

$$T^{-1} \sum \tilde{y}_{t-1} \hat{w}_t = T^{-1} \sum \tilde{y}_{t-1} \left[h(\varepsilon_t) + (\hat{\varepsilon}_t - \varepsilon_t) h'(\varepsilon_t) - \hat{\varepsilon}_t \hat{D}_2 \right] + o_p(1)$$

Therefore,

$$\begin{aligned} & T^{-1} \sum \tilde{y}_{t-1} (\hat{w}_t - w_t) & (A.21) \\ = & T^{-1} \sum \tilde{y}_{t-1} \left[(\hat{\varepsilon}_t - \varepsilon_t) h'(\varepsilon_t) - (\hat{\varepsilon}_t - \varepsilon_t) \hat{D}_2 - \varepsilon_t (\hat{D}_2 - D_2) \right] + o_p(1) \end{aligned}$$

But,

$$T^{-1} \sum \tilde{y}_{t-1} (\hat{\varepsilon}_t - \varepsilon_t) h'(\varepsilon_t) = T (\hat{\beta} - \beta) T^{-2} \sum \tilde{y}_{t-1}^2 h'(\varepsilon_t) + o_p(1), \quad (A.22)$$

$$T^{-1} \sum \tilde{y}_{t-1} (\hat{\varepsilon}_t - \varepsilon_t) \hat{D}_2 = \hat{D}_2 T (\hat{\beta} - \beta) T^{-2} \sum \tilde{y}_{t-1}^2 + o_p(1), \quad (A.23)$$

and

$$T^{-1} \sum \tilde{y}_{t-1} \varepsilon_t (\hat{D}_2 - D_2) = o_p(1). \quad (A.24)$$

Two terms (A.22) and (A.23) cancel each other in the limit in (A.21), so the proof is complete.

References

- [1] Amemiya, T. (1985), *Advanced Econometrics*, Basil Blackwell Ltd.
- [2] Beelder, O. (1996), "Adaptive estimation and unit root tests," manuscript, Department of Economics, Rochester University.
- [3] Cox, D.D. and I. Llatas (1991), "Maximum likelihood type estimation for nearly nonstationary autoregressive time series," *Annals of Statistics*, 19, 1109-1128.
- [4] Hansen, B. E. (1995), "Rethinking the univariate approach to the unit root testing: using covariates to increase the power," *Econometric Theory*, 11, 1148-1171.
- [5] Hasan, M.N. and R.W. Koenker (1997), "Robust rank test of unit root hypothesis," *Econometrica*, 65, 133-161.
- [6] Herce, M.A. (1996), "Asymptotic theory of LAD estimation in a unit root process with finite variance errors," *Econometric Theory*, 12, 129-153.
- [7] Im K.S. and P. Schmidt (1999), "More efficient estimation under non-normality when higher moments do not depend on the regressors, using residual-augmented least squares," manuscript, Department of Economics, Wichita State University.
- [8] Knight, K. (1989), "Limit theory for autoregressive-parameter estimates in an infinite-variance random walk," *The Canadian Journal of Statistics*, 17, 261-278.
- [9] Lucas, A. (1995), "Unit root tests based on M-estimators," *Econometric Theory*, 11, 331-346.