UPPSALA
UNIVERSITET

Working Paper 2013:8

Department of Statistics

The Number of Bootstrap Replicates in Bootstrap Dickey-Fuller Unit Root Tests

Jianxin Wei

Working Paper 2013:8
June 2013
Department of Statistics
Uppsala University
Box 513
SE-751 20 UPPSALA
SWEDEN

Working papers can be downloaded from www.statistics.uu.se

Title: The Number of Bootstrap Replicates in Bootstrap Dickey-Fuller Unit Root Tests
Author: Jianxin Wei
E-mail: jianxin.wei@statistics.uu.se

The Number of Bootstrap Replicates in Bootstrap Dickey-Fuller Unit Root Test

Jianxin Wei

Abstract

The Dickey-Fuller unit root test, using asymptotic critical value, exhibits size distortion and the bootstrap method, if used correctly, can overcome this problem. This study focuses on the question how many bootstrap replicates B are necessary. Through a simulation study, we found that a small B ($B = 19$) is enough for a precise size. However, with a too small B we will lose power when the null hypothesis is not true.

1 Introduction

In unit root testing, when using e.g. the Dickey-Fuller test, the small sample distribution of the test statistic is unknown and the asymptotic distribution is instead used for making inference. However, when the sample size is small, the discrepancy between the actual and asymptotic distribution is large and consequently it will affect the size of the test. The bootstrap method, if used correctly, can provide asymptotic refinement, i.e. the bootstrap distribution can approximate the actual distribution up to the order $O_p(T^{-\frac{1}{2}})$ and improve the size, see Park (2003). In practice, a feasible bootstrap test must use a finite number of bootstrap replicates and, as discussed in Davidson and MacKinnon (2001), there will be mainly two undesirable consequences of using a finite B. First, the result of a test may depend on the sequence of random numbers used to generate the bootstrap samples. Second, there will be some loss of power, see among others, Hall and Titterington (1989). A practical question is how to choose the number of bootstrap replicates B. Theoretically, the larger B the better. But if we consider the computational cost of running the bootstrap program, especially in a simulation study, we

want to find a small B without losing any essential properties. In this paper, the number of bootstrap replicates B is studied from two perspectives. Firstly, we consider the size of the test, i.e. how large B is enough to obtain a nice size property. Secondly, we also consider the power. Since small B may cause power loss, see, among others Hall and Titterington (1989), we should increase B to get larger power even if a small B is enough to get a nice size property. The paper is organized as follows, in Section 2, DF unit root test and the bootstrap method is reviewed. Section 3 presents the design of the experiment. Section 4 discusses the simulation result.

2 Dickey-Fuller unit root tests and the Bootstrap

Consider an autoregressive (AR) model of order 1,

$$Y_t = \rho Y_{t-1} + \varepsilon_t, \qquad t = 1, 2, ..., T.$$

We are interested in testing $H_0 : \rho = 1$, i.e. Y_t has a unit root, against $H_A : \rho < 1$, i.e. Y_t is stationary. The parameter ρ is estimated by the ordinary least square (OLS) estimator,

$$\hat{\rho} = \frac{\sum_{t=1}^{T} y_t y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2}.$$

The two test statistics based on $\hat{\rho}$ are the t test statistic

$$F_T = \frac{(\hat{\rho} - 1)}{se(\hat{\rho})},$$

where $se(\hat{\rho}) = (\frac{\sum_{t=1}^{T}(y_t - \hat{\rho}_T y_{t-1})^2}{T-1} / \sum_{t=1}^{T} y_{t-1}^2)^{\frac{1}{2}}$ is the standard error of $\hat{\rho}$ and the coefficient test statistic

$$G_T = T(\hat{\rho} - 1).$$

The small sample distributions of F_T and G_T are unknown and they have the Dickey-Fuller type of asymptotic distribution:

$$F_T \xrightarrow{d} \frac{\frac{1}{2}[W^2(1) - 1]}{\{\int_0^1 W^2(r)dr\}^{\frac{1}{2}}},$$

$$G_T \xrightarrow{d} \frac{\frac{1}{2}[W^2(1) - 1]}{\int_0^1 W^2(r)dr}.$$

When using critical value from the asymptotic distribution there is size distortion and sometimes the discrepancy is too large for the tests to be reliable. The bootstrap unit root test is an attractive approach to reduce the discrepancy of the empirical size and the nominal size. Park (2003) gave a theoretical proof why the bootstrap unit root test has a better size property than using the asymptotic critical values. He proved that the bootstrap gives asymptotic refinement in the unit root test under the assumptions (a) $E\varepsilon^r < \infty$ for some $r > 12$ and (b) the AR order p is known. The result from Park (2003) is applicable to a wide class of models including the model with short run dynamics and the model with different deterministic terms. The implementation of the bootstrap is straightforward. First we fit the model under the null hypothesis and obtain the residuals $\{\hat{\varepsilon}_t\}$. For the model with no short run dynamics, this is trivial since $\hat{\varepsilon}_t = y_t - y_{t-1}$. Second, we apply i.i.d. bootstrap to the re-centered residuals $\{\hat{\varepsilon}_t - \bar{\hat{\varepsilon}}_\cdot\}$[1] to obtain the bootstrap residuals $\{\varepsilon_t^*\}$. In the end, construct the bootstrap sample y_t^* under the null using the bootstrap residuals. We construct the bootstrap samples y_t^* starting from $y_1^* = y_1$ and calculate the bootstrap statistics of our interest F_T^* or G_T^*. After repeating the procedure B times, we have B bootstrap statistics and the bootstrap P-value is estimated by the proportion of bootstrap samples that yield a statistic smaller than F_T,

$$\hat{p}^*(F_T) = \frac{1}{B+1} \sum_{j=1}^{B} I(F_T^* < F_T),$$

where $I(\cdot)$ is the indicator function. The bootstrap test rejects the null hypothesis if $\hat{p}^*$ is smaller than the nominal size α. Note that $\hat{p}^*$ is only an estimated P-value since the true bootstrap P-value p^* is from the ideal bootstrap test, i.e. when $B \to \infty$. As $B \to \infty$, the estimated P-value $\hat{p}^*$ will tend to the ideal P-value p^*. In practice, one can choose extremely large B to ensure that the difference of $\hat{p}^*$ and p^* can be neglected. However, sometimes it is not cheap to compute the bootstrap statistic or, in some simulation studies, an extremely large number of replications is needed and the overall computational cost is large. In such cases, we want B to be fairly small. In next section, we will conduct an Monte Carlo experiment to study the number of bootstrap replicates B in Dickey-Fuller unit root test.

[1] $\bar{\hat{\varepsilon}}_\cdot = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_t$. If intercept terms are included in the regression, the recentering is unnecessary because the sum of the residuals is zero

3 Monte Carlo Simulation

The Monte Carlo simulation study is conducted from two perspectives, size and power. To study the size, the data generating process (DGP) is under the null hypothesis i.e. $Y_t = Y_{t-1} + \varepsilon_t$. Since both the two test statistics are asymptotically pivotal, for doing an exact test, the number of bootstrap replicates B must satisfy that $\alpha(B+1)$ is an integer, where α is the significance level, see Dufour and Kiviet (1998) for details. This is crucial when using a small B. However, the effect is negligible if B is large. In our study, we consider the most common significance level $\alpha = 0.05$, therefore the possible values for B are 19, 39, etc. B is chosen as $B = 19, 39, 99$ at the initial stage. In the simulation we found that the empirical rejection probability is satisfactory, since it is close to the nominal level and therefore it is not necessary to use a larger B. We consider two different error terms in the simulation, standard normal $N(0,1)$ and demeaned chi-square $\chi^2(8) - 8$(skewed), where both have zero mean. The results are shown in Table 1 and Table 2.

For the power study, the DGP is $Y_t = 0.9Y_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim N(0,1)$. The simulation is also done with $B = 19, 39, 99$ at first and we found that the power increases as B increases (as expected) especially for large sample sizes ($T = 50, 100$). This suggests that there could be space for the power to increase if we continue to increase B, so we did the simulation again by using $B = 19, 39, 99, 199, 399$. The result is shown in Table 3 together with the raw power and the size adjusted power of the asymptotic test.

We also did a simulation for the model with an intercept term (corresponding to Case 2 in Chapter 17 in Hamilton (1994)). In this case, the true model is random walk $Y_t = Y_{t-1} + \varepsilon_t$, and the data is fitted to the model $Y_t = \alpha + \rho Y_{t-1} + \varepsilon_t$. The result is shown in Table 5 and Table 6. In all studies, the asymptotic critical values are from Table B5 and B6 in Hamilton (1994), the sample size $T \in \{10, 25, 50, 100\}$, the number of replicates is 1,000,000.

4 Results and concluding remarks

As can be seen from Table 1, Table 2 and Table 5 the empirical size (rounded off to one decimal) remains almost the same as B increases. This suggests $B = 19$ is satisfactory for the study on the empirical size. This is surprising because 19 observations is not dense enough for an empirical distribution to mimic the whole distribution under the null, in particular the tail properties.

In addition, many empirical rejection probabilities in Table 1 and Table 2 are 5 percent when rounding off to one decimal. These perfect results may be because all our simulations are based on the simple AR(1) model and these perfect rejection probabilities may not be obtained if the model includes short run dynamics and deterministic terms. For example, in Table 5, when an intercept is included in the model, the estimated size of the bootstrap test is 5.4 percent when $B = 19$ and exhibits slightly change when $B = 39, 99$.

While looking at the power, when the sample size is small, say, T smaller than 25, the power loss from $B = 19$ to $B = 399$ is quite small. For example, in Table 3, F_T, $T = 10$, the power of $B = 19$ is 7.8 percent and the power of $B = 399$ is 7.9 percent, the difference is negligible. The very small power is because of the inherent property of the unit root test which could not be overcome by the bootstrap. When the sample size is small, the power is small even if the actual critical value is used. Since the bootstrap is a way to approximate the actual distribution, it cannot necessarily provide larger power than the asymptotic test. For medium sample size, $T = 50$, the discrepancy of the power by using different B is relatively large and for a large sample size $T = 100$ the discrepancy is larger. For example, in Table 3, F_T, $T = 100$, the power of $B = 19$ is 62.5 percent whereas the power of $B = 399$ is 76.1 percent. In these cases the distribution under the null and the alternative are more distinguishable such that too small B is not enough to distinguish the bootstrap null distribution from the distribution under the alternative. The power properties for the χ^2 error terms, as shown in Table 4, are similar to the power property of the normal error term.

The relation between the power of the test using the asymptotic critical value and the bootstrap test are different for F_t and G_T. For F_T, (corresponding to the t test), the power of the asymptotic test is larger than the power of the bootstrap test. This is because of the fact that the distribution of F_T will shift to the right as the sample size increases and therefore the asymptotic critical value is larger than the actual critical value and the bootstrap critical value (since the bootstrap is to approximate the actual distribution). This is in contrast with G_T (corresponding to the coefficient test). The distribution of G_T shifts to the left as the sample size increases and the asymptotic critical value is smaller than the actual critical value and in consequence the test using the asymptotic critical value has smaller power than the bootstrap test. For both F_T and G_T, the powers of the bootstrap test and the asymptotic test become close as the sample size increases.

For the size adjusted power, the bootstrap power[2] is higher than the size adjusted power for both F_T and G_T except for $T = 100$. This is probably because, for $T = 100$, there is more space to increase the power if we increase B.

Overall the bootstrap has better performance in both size and power, however, the difference is small when the sample size is large. A small number of B is enough if we focus on analyzing the size property of the bootstrap test but not the power.

References

Davidson, R. and MacKinnon, J. G. (2001). Bootstrap tests: How many bootstraps?, *Technical Report 1036*, Queen's University, Department of Economics.

Dufour, J.-M. and Kiviet, J. F. (1998). Exact inference methods for first-order autoregressive distributed lag models, *Econometrica* **66**: 79–104.

Hall, P. and Titterington, D. M. (1989). The effect of simulation order on level accuracy and power of monte carlo tests, *Journal of the Royal Statistical Society. Series B (Methodological)* **51**: 459–467.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press.

Park, J. Y. (2003). Bootstrap unit root tests, *Econometrica* **71**: 1845–1895.

Tables

[2]Since the definition of the bootstrap size adjusted power is not clear, we use the bootstrap power to compare with the size adjusted power.

Table 1: Empirical size in percent, $\varepsilon_t \sim N(0,1)$.

	F_T			
	B=19	B=39	B=99	Asy
T=10	5.0	5.0	5.0	6.2
T=25	5.0	5.0	5.0	5.4
T=50	5.0	5.0	5.0	5.2
T=100	5.0	5.0	5.0	5.1
	G_T			
	B=19	B=39	B=99	Asy
T=10	5.0	5.0	5.0	2.5
T=25	5.0	5.0	5.0	3.9
T=50	5.0	5.0	5.0	4.4
T=100	5.0	5.0	5.0	4.6

Table 2: Empirical size in percent, $\varepsilon_t \sim \chi^2(8) - 8$.

	F_T			
	B=19	B=39	B=99	Asy
T=10	4.9	4.9	4.9	5.9
T=25	5.0	5.0	4.9	5.3
T=50	5.0	5.0	5.0	5.1
T=100	5.0	5.0	5.0	5.0
	G_T			
	B=19	B=39	B=99	Asy
T=10	4.9	5.0	4.9	2.4
T=25	5.0	5.0	4.9	3.8
T=50	5.0	5.0	5.0	4.3
T=100	5.0	5.0	5.0	4.6

Table 3: Empirical power in percent, $\varepsilon_t \sim N(0,1)$. Asy(raw) represents the raw power, Asy(Ad.) represents the size adjusted power.

			F_T				
	B=19	B=39	B=99	B=199	B=399	Asy(raw)	Asy(Ad.)
T=10	7.8	7.8	7.9	7.9	7.9	9.8	5.5
T=25	14.1	14.4	14.6	14.7	14.7	16.0	12.2
T=50	29.0	30.7	32.0	32.4	32.6	33.8	31.8
T=100	62.5	68.9	73.6	75.3	76.1	77.3	77.5
			G_T				
	B=19	B=39	B=99	B=199	B=399	Asy(raw)	Asy(Ad.)
T=10	7.7	7.7	7.8	7.8	7.8	3.9	5.2
T=25	13.9	14.2	14.4	14.5	14.5	11.5	12.4
T=50	28.7	30.3	31.5	31.9	32.1	29.2	31.1
T=100	62.4	68.7	73.4	75.1	76.0	74.8	77.1

Table 4: Empirical power in percent, $\varepsilon_t \sim \chi^2(8) - 8$. Asy(raw) represents the raw power, Asy(Ad.) represents the size adjusted power.

			F_T				
	B=19	B=39	B=99	B=199	B=399	Asy(raw)	Asy(Ad.)
T=10	7.5	7.6	7.6	7.6	7.7	9.5	5.8
T=25	13.8	13.9	14.2	14.4	14.5	15.5	11.6
T=50	28.0	28.5	30.5	31.8	32.2	33.2	31.0
T=100	60.2	67.9	71.9	73.9	75.0	76.1	77.2
			G_T				
	B=19	B=39	B=99	B=199	B=399	Asy(raw)	Asy(Ad.)
T=10	7.4	7.4	7.5	7.6	7.6	3.7	6.0
T=25	13.5	13.9	14.0	14.0	14.0	11.1	14.3
T=50	25.5	27.3	30.5	31.1	32.1	29.5	32.0
T=100	61.1	64.7	70.4	73.1	75.7	73.8	76.0

Table 5: Empirical size in percent, model with intercept, $\varepsilon_t \sim N(0,1)$.

	F_T			
	B=19	B=39	B=99	Asy
T=10	5.4	5.5	5.5	8.6
T=25	5.1	5.1	5.1	6.4
T=50	5.0	5.0	5.0	5.7
T=100	5.0	5.0	5.0	5.4
	G_T			
	B=19	B=39	B=99	Asy
T=10	5.0	5.1	5.1	0.4
T=25	5.0	5.0	5.0	2.5
T=50	5.0	5.0	5.0	3.7
T=100	5.0	5.0	5.0	4.3

Table 6: Empirical power in percent, model with intercept, $\varepsilon_t \sim N(0,1)$. Asy(raw) represents the raw power, Asy(Ad.) represents the size adjusted power.

	F_T					
	B=19	B=99	B=199	B=399	Asy(raw)	Asy(Ad.)
T=10	6.0	6.1	6.1	6.1	9.5	7.2
T=25	7.0	7.1	7.1	7.1	9.0	8.0
T=50	11.2	11.6	11.6	11.7	13.3	10.6
T=100	27.5	30.4	30.9	31.1	33.0	28.0
	G_T					
	B=19	B=99	B=199	B=399	Asy	Asy(Ad.)
T=10	7.3	7.4	7.4	7.4	0.6	3.2
T=25	10.7	11.2	11.2	11.3	6.0	11.6
T=50	18.0	19.4	19.6	19.7	15.4	24.2
T=100	40.4	46.0	46.9	47.3	43.8	52.0