Bootstrapping the Breusch-Godfrey autocorrelation test for a single equation dynamic model: Bootstrapping the Restricted vs. Unrestricted model

Panagiotis Mantalos
Department of Health, Science, and Mathematics
Blekinge Institute of Technology
Sweden

ABSTRACT

We use Monte Carlo methods to study the properties of the bootstrap Breusch-Godfrey test for autocorrelated errors in two versions a) by bootstrapping under the null hypothesis, restricted and b) by bootstrapping under the alternative hypothesis, unrestricted. We use the residual bootstrap for the bootstrap-BG test. Our analysis regarding the size of the test reveals that both bootstrap tests have actual sizes that lie close to the nominal size, with the restricted being better. Regarding the power of the test we find that with bootstrapping under the alternative hypothesis, the unrestricted bootstrap test has the greater power in small samples.

Keywords: Autocorrelation, Bootstrap, Breusch-Godfrey test method
JEL Classification Codes: C12, C15

1. Introduction

The history of autocorrelation tests dates back to the paper by Durbin and Watson (1950), who introduced their now classic test for autocorrelated errors in a regression model. However, the Durbin-Watson (D-W) statistic tests only for autocorrelation of the first order, and it is not valid in dynamic models (Maddala, 1995).

One alternative that has been suggested is to use the Breusch-Godfrey (BG) test. This test, introduced by Breusch (1978) and Godfrey (1978), is easy to apply, applicable in the presence of lagged dependent variables, valid for very general hypotheses about the serial correlation in the errors, and is asymptotically equivalent to the Lagrange Multiplier (LM) test. Kiviet (1986) used Monte Carlo methods to compare different LM, Wald and LR alternatives for dynamic single equation models, and showed that using standard *F*-tests in the second equation was to be preferred. Edgerton and Shukur (1999), who studied the properties of various generalizations of the test in a system perspective, also found that using a system-wise Rao's *F*-test (proposed by Rao, 1973) leads to superior properties when testing for autocorrelation.

In both the last named papers and generally when we study the small sample properties of a test procedure by comparing different tests, two aspects are of prime importance:

a) finding the test that has actual size closest to the nominal size, and given that (a) holds;
b) finding the test that has the greatest power.

In most cases, however, the distributions of the test statistic we use are known only asymptotically, and unfortunately, unless the sample size is very large indeed, the tests may not have the correct size and inferential comparisons and judgements based on them might be misleading.

One way to deal with this situation that has emerged in recent years is to use the bootstrap technique. However, using this technique for testing autocorrelation is not new. Davidson and MacKinnon (1996) applied it to test for autocorrelation of the first order in a dynamic model. Morey and Wang (1985) bootstrapped the Durbin-Watson test as did Jeong and Chung (2001). These three studies all show the superiority of the bootstrap technique in finite samples.

The issue of bootstrapping a test statistic, even if it is well studied, is not trivial. One of the basic problems is deciding how to resample the data, and whether to resample under the null hypothesis or under the alternative hypothesis. In this paper, by using the Monte Carlo procedure we will investigate the properties of the bootstrap-BG test procedure in two different cases, first by bootstrapping under the null hypothesis (that is, bootstrapping the restricted model) and bootstrapping under the alternative hypothesis (that is, bootstrapping the unrestricted model). We will use mainly the Residual Bootstrap (RB) to study the properties of the BG test procedure when the errors are IID.

The rest of the paper is organized as follows: Section 2 presents a review of the BG-test. In Section 3 we introduce the data, the model and the methodology. Section 4 presents the estimated results concerning the size of the tests, while Section 5 presents the estimated results concerning the power of the tests. Finally, we give a short summary and conclusion in Section 6.

2. **Breusch-Godfrey Test**

Consider the general single equation dynamic model

$$\begin{aligned} y_t &= a + \mathbf{X}_t\beta + \gamma y_{t-1} + e_t, \\ e_t &= \rho e_{t-1} + u_t \qquad u_t \sim NID(0,\sigma^2) \end{aligned} \tag{1}$$

where y_t and e_t are $(T\times 1)$, $\boldsymbol{X}_t$ is $(T\times m)$, β is $(m\times 1)$ and T denotes the number of observations.

Equation (1) is called the primary regression. The BG test is performed by first calculating the least squares residuals $\hat{\mathrm{e}} = (\mathbf{I}_T - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{Y}$ from this regression, where $\boldsymbol{Y}$ is the $(T\times 1)$ matrix of endogenous variables, and $\mathbf{Z} = (\mathbf{I}_T\mathbf{X}\ y_{-1})$ is the $(T\times(m+2))$ matrix of exogenous, constant and lagged endogenous variable. These residuals are then used in the following auxiliary equation:

$$\mathrm{y}_t = A + \mathbf{X}_t\mathrm{B} + \Gamma_1 y_{t-1} + \psi_1\hat{\mathrm{e}}_{t-1} + \ldots + \psi_G\hat{\mathrm{e}}_{t-G} + \delta_t . \tag{2}$$

The BG test is now performed by testing the hypothesis $H_0 : \psi_1 = \cdots = \psi_G = 0$.

Now we denote by $\hat{\delta}_U$ the vector of estimated residuals from the *unrestricted* regression (2), and by $\hat{\delta}_R$ the equivalent vector of residuals from the *restricted* regression with H_0 imposed. Defining the matrix of cross-products of these residuals as

$$\mathbf{S}_U = \hat{\delta}_U' \hat{\delta}_U \tag{3a}$$

and

$$\mathbf{S}_R = \hat{\delta}_R' \hat{\delta}_R , \tag{3b}$$

the Wald, Likelihood Ratio and Lagrange Multiplier test statistics are given by

$$W = \tau(\mathrm{tr}\,\mathbf{S}_U^{-1}\mathbf{S}_R - 1), \tag{4a}$$

$$LR = \tau \ln U , \text{ and} \tag{4b}$$

$$LM = \tau(1 - \mathrm{tr}\,\mathbf{S}_R^{-1}\mathbf{S}_U), \tag{4c}$$

where $U = \det \mathbf{S}_R / \det \mathbf{S}_U$ and τ is the rows of $\hat{\mathrm{e}}$. The above statistics are all asymptotically $\chi^2(p)$ distributed under the null hypothesis, where p is the number of restrictions imposed by H_0.

Note again that all the above statistics have been defined from the auxiliary regression (2), estimated using T observations.

The F test statistic has the usual form:

$$F = \frac{(\mathbf{S}_R - \mathbf{S}_U)/p}{\mathbf{S}_U/(T-K)} \tag{4d}$$

3. The Monte Carlo Experiment

The Monte Carlo experiment was performed by generating data according to (1), estimating the auxiliary regression (2), and then calculating the test statistics 4a-4d defined in Section 2. For simplicity and without loss of generality, we use just one exogenous variable and one lagged residual in equation (2). That is, we estimate the auxiliary equation:

$$\mathrm{y}_t = A + BX_t + \Gamma_1 y_{t-1} + \psi_1 \hat{\mathrm{e}}_{t-1} + \delta_t \tag{2a}$$

The u_t errors of (1) are NID(0,1).

For each model we performed 10,000 replications for the calculation of the sizes, and 1000 for the power functions.
The distribution of the exogenous variables that we use is the following fairly general type of simple AR(1) generating process:

$$X_t = \phi X_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim NID(0, \sigma^2) \tag{5}$$

The parameter of the exogenous AR process ϕ is 0.50, the same for the dynamic parameter γ = 0.50, and the numbers of observations are 25, 50 and 100.

3.1 The P-value plots

The conventional way to report the results of a Monte Carlo experiment is to tabulate the proportion of how many times the null hypothesis is rejected in repeated samples under conditions where the null is true.

Concerning the significance levels to be used when judging the properties of the tests, different authors have put forward reasons for using both larger and smaller significance levels. Maddala (1992) suggests using significance levels of as much as 25% in diagnostic testing, while MacKinnon (1992) suggests going in the other direction.

To reduce this problem, in this study we use mainly graphical methods that may provide more information about the size and the power of the tests. We use the simple graphical methods developed and illustrated by Davidson and MacKinnon (1997) that are easy to interpret, the "P-*value plot*" to study the size, and the "*Size-Power curves*" to study the power of the tests. The graphs, the P-value plots and Size-Power curves are based on the empirical distribution function, EDF of the P-values, denoted as $\hat{F}\left(x_j\right)$.

3.2 Bootstrap-critical values, bootstrapping the restricted model

As mentioned in the introduction, in most cases the distributions of the test statistics that we use are known only asymptotically. As a result, the tests may not have the correct size, and inferential comparisons and judgments based on them might be misleading. However, by

using bootstrap technique we can improve the critical values so that the true size of the test approaches its nominal value.

It is often assumed that the errors are normally distributed, but for our experiment, even if we have simulated error terms that are NID(0,1), we assume only that the errors are IID. In that case the simplest method is to bootstrap the residuals. Suppose the original estimation yields residuals $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_T$. Then we obtain bootstrap error terms by resampling with replacement from the leverage-adjusted residuals:

$$\left(\frac{T}{T-k}\right)^{1/2}\left(\frac{\hat{e}_t}{(1-h_t)^{1/2}} - \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{e}_t}{(1-h_t)^{1/2}}\right) \tag{6}$$

where $h_t = Z_t'\left(\mathbf{Z}'\mathbf{Z}\right)^{-1} Z_t$. Each element of each vector of bootstrap errors is one of the leverage-adjusted residuals, chosen at random with probability 1/T.

As early as Freedman's (1984) paper, if the model is dynamic, it is customary to generate the y_t^* recursively. That is, we generate the bootstrap samples as follows:

(1) We estimate the test statistic as described in Section 2a, (4), which we call T_s.

(2) We use the adjusted OLS residuals $\left(\frac{T}{T-k}\right)^{1/2}\left(\frac{\hat{\delta}_t}{(1-h_t)^{1/2}} - \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\delta}_t}{(1-h_t)^{1/2}}\right)$ i = 1,...,T. to draw i.i.d $\delta_1^*, \ldots, \delta_T^*$ data and generate the y_t^* recursively by defining:

$$y_t^* = \hat{A} + \hat{B}X_t + \hat{\Gamma}_1 y_{t-1}^* + \delta_t^*.$$

(3) We then calculate the test statistics T_s^* as described in Section 2, equations 4a-4d.

(4) Repeating this step N_b times and taking the (1-α):*th* quintile of the bootstrap distribution of T_s^*, we obtain the α - level "bootstrap critical values" ($c_{t\alpha}^*$), and, finally, we then reject Ho if $T_s \geq c_{t\alpha}^*$.

Among papers that advocate this approach are those of Horowitz (1994), Shukur and Mantalos (1997a) and Mantalos and Shukur (1998), whereas Davidson and MacKinnon (1996) advocate the estimate of the P-value.

A bootstrap estimate of the P-value for testing is P*{ $T_s^* \geq T_s$ }, and this is the approach that we use to study the size of the bootstrap test with the assistance of the "P-*value plot*". As for N_b, which is the size of the bootstrap sample used to estimate bootstrap critical values, Horowitz (1994) uses N_b =100, while Davidson and Mackinnon (1996) use N_b =1000 to estimate the P-value. In our study we use N_b = 200.

3.3 Bootstrap-hypothesis testing, bootstrapping the unrestricted model

One of the important considerations for generating the y_t^* is to impose the null hypothesis on the model from which we generate the y_t^*. However, some authors, including Jeong and Chung (2001), argue for bootstrapping under the alternative hypothesis. "Let the data speak" is their basic bootstrap principle. In our case we resample the data as follows:

(1) We calculate first the least squares residuals $\hat{e}_t$ from the primary regression. These residuals are then used in the following auxiliary equation:

$$y_t = A + BX_t + \Gamma_1 y_{t-1} + \psi_1 \hat{e}_{t-1} + \delta_t$$

to estimate the coefficient $\hat{\psi}_1$ and hence the studentized "pivot" $T_s = (\hat{\psi}_1)/\hat{\sigma}_T$, where $\hat{\sigma}_T^2$ is the estimate variance of $\hat{\psi}_1$.

(2) We use the adjusted OLS residuals $\left(\frac{T}{T-k}\right)^{1/2}\left(\frac{\hat{\delta}_t}{(1-h_t)^{1/2}} - \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\delta}_t}{(1-h_t)^{1/2}}\right)$ i = 1,...,T. to draw i.i.d $\delta_1^*,\ldots,\ \delta_T^*$ data and generate the y_t^* by generating first the e_t^* recursively:

$$\hat{e}_t^* = \hat{\psi}_1 \hat{e}_{t-1}^* + \delta_t^*$$

and then

$$y_t^* = \hat{A} + \hat{B}X_t + \hat{\Gamma}_1 y_{t-1}^* + e_t^*.$$

Note that the first observations for the e_t^* and y_t^* are: $e_1^* = \frac{\delta_1^*}{\sqrt{1-\hat{\psi}_1^2}}$ and $y_1^* = \frac{e_1^*}{\sqrt{1-\hat{\Gamma}_1^2}}$

(3) We then repeat step (1) to calculate the bootstrap coefficient $\hat{\psi}_1^*$ and then the $T_s^* = (\hat{\psi}_1^* - \hat{\psi}_1)/\hat{\sigma}_T^*$, where $(\hat{\sigma}_T^*)^2$ is the estimated- variance of $\hat{\psi}_1^*$.

(4) Repeating this step N_b times we estimate that the P-value for testing is P*{ $T_s^* \geq T_s$ }, and this is the approach that we use to study the size of the bootstrap test with the assistance of the "P- *value plot*". The N_b here is 200.

4. Analysis of the Size

In this section we present the results of our Monte Carlo experiment concerning the size of the bootstrap tests.

For the P-value plots, if the distribution used to compute the p_s is correct, each of the p_s should be distributed uniformly on (0,1). Therefore the resulting graph should be close to the 45^o line. Furthermore, to judge the reasonableness of the results we use a 95% confidence interval for the actual size (π_0) as : $\pi_0 \pm 2\sqrt{\frac{\pi_0(1-\pi_0)}{N}}$, where N is the number of Monte Carlo replications. Results that lie between these bounds will be considered satisfactory. For example, if we consider a nominal size of 5%, we define a result as reasonable if the estimated size lies between 4.46% and 5.44%.

The P-value plots also make it possible and easy to distinguish between tests that systematically over-reject or under-reject, and tests that reject the null hypothesis about the right proportion of the time.

Figure 1 shows the truncated P-value plots for the actual size of the bootstrap, the F and the Wald tests, using 25, 50 and 100 observations. Looking at these curves, it is not difficult to make the inference that the bootstrap test performs adequately, as it lies inside the confidence bounds. The same holds for the F-test, as it also lies inside the confidence bounds. However, using the asymptotic critical values, the Wald test seems to show a slight tendency to over-reject the null hypothesis.

Figure 1 P-value plot, Parametric vs. Bootstrap Tests

Figure 1a 25 Observations

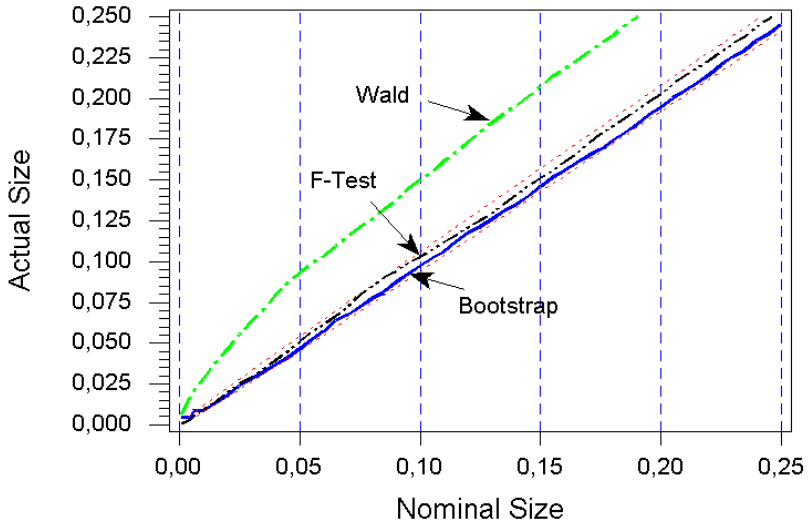


Figure 1b 50 Observations

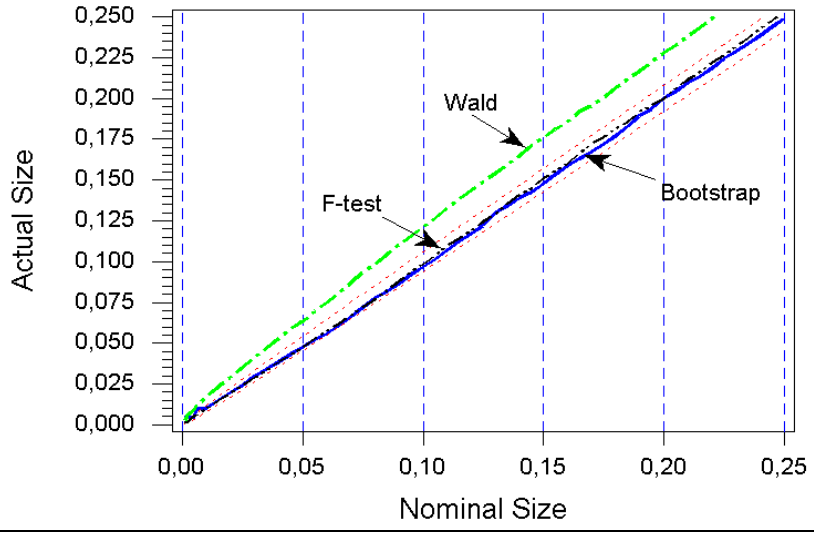
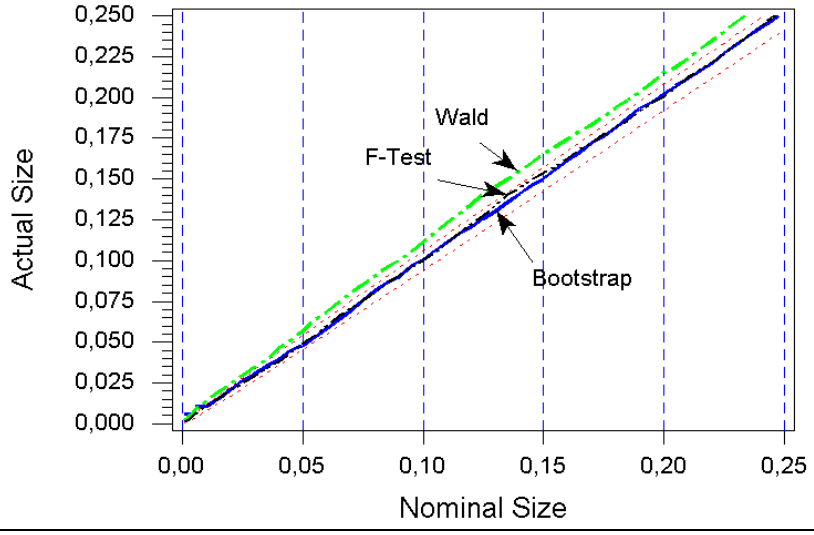


Figure 1c 100 Observations



Dot line: Confidence interval

Figure 2 P-value plot, Restricted vs. Unrestricted Bootstrap Tests

Figure 2a 25 Observations

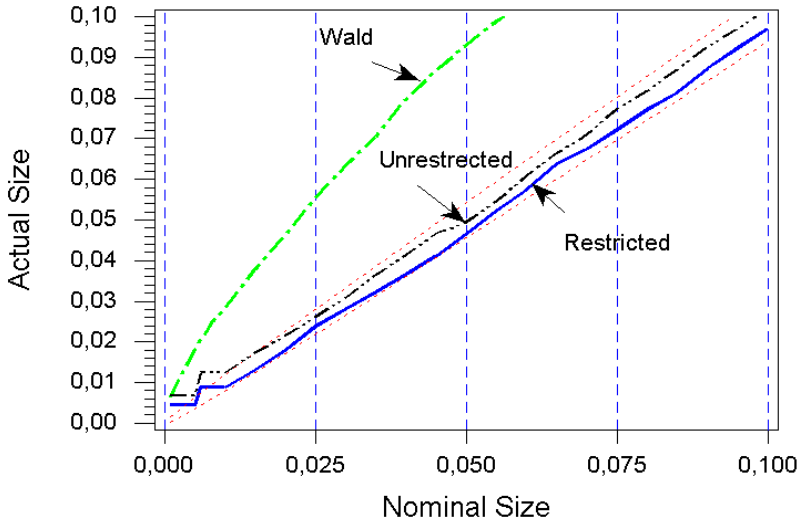


Figure 2b 50 Observations

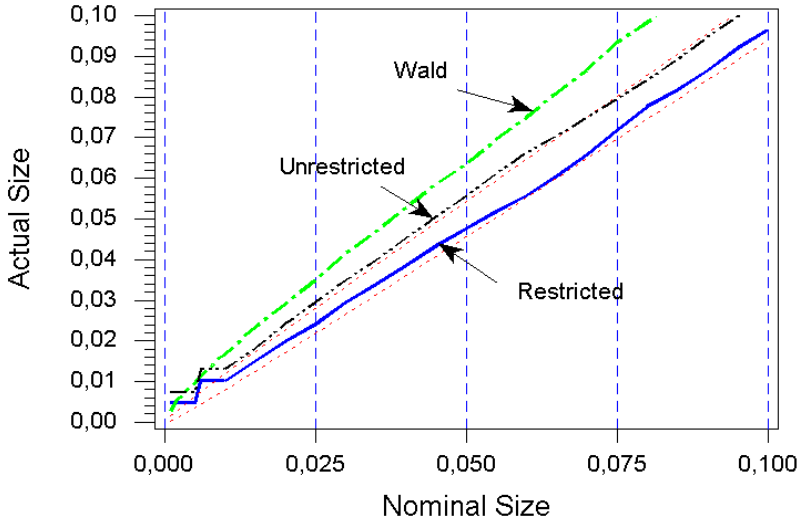
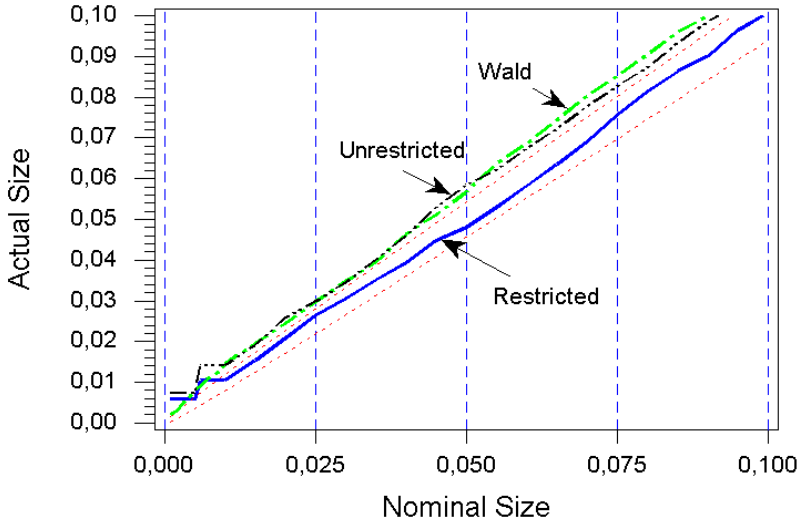


Figure 2c 100 Observations



Dot line: Confidence interval

The superiority of the bootstrap test, concerning the size of the tests, is for most F-Tests negligible but for most Wald tests considerable, and this is more noticeable in small samples, 25 and 50.

Figure 2 shows the truncated P-value plots up to 10% for the actual size of the restricted and unrestricted bootstrap tests. We truncated up to 10% to facilitate the comparison between two well-performed bootstrap tests. However, as we see, there is a small difference between the restricted and unrestricted bootstrap tests at the significance levels of as much as 5%, the unrestricted bootstrap lying on the upper bounds or just outside the bound for 100 observations, with the restricted lying inside the bounds. Note also that all four bootstrap tests (Wald, LRE, LM and F-test) have identical results in the restricted case, while in the unrestricted case we use only the studentized "pivot" bootstrap-t test.

To summarize the results concerning the size of the test, we find that the bootstrap tests perform adequately in all samples, while the restricted bootstrap is just a little better than the unrestricted bootstrap method in large samples.

5. Analysis of the Power of the Tests

In this section we analyse the power of the Wald and bootstrap tests using sample sizes of 25, 50 and 100 observations. The power function is estimated by calculating the rejection frequencies in 1000 replications using the value $\rho = 0.4$.

We used the Size-Power Curves to compare the estimated power functions of the alternative test statistics. This proved to be quite adequate, because those tests that gave reasonable results regarding size usually differed very little regarding power. We followed the same process as for the size investigation (see Section 4) to evaluate the EDFs denoted by $\hat{F}^{\oplus}\left(x_j\right)$, by using the same sequence of random numbers as that which we used to estimate the size of the tests.

Using Size-Power Curves to plot the estimated power functions against the nominal size. While plotting the estimated power functions against the true size, that is $\hat{F}^{\oplus}\left(x_j\right)$ against

$\hat{F}\left(x_j\right)$, we have the Size-Power Curves on a correct size-adjusted basis. Figure 3 shows the results of using the Size-Power Curves. We see that the unrestricted bootstrap test method is now superior not only against the Wald test but also against the restricted bootstrap. The most interesting result is that the superiority is much more noticeable in small samples. We also see a sample effect: the larger the sample, the larger is the power of the tests. Finally, as the sample size increases, the power difference decreases, showing that the Wald test has higher power than the restricted bootstrap.

However, when using the Size-Power Curves on a correct size-adjusted basis, the situation is different concerning the power of the Wald and the restricted bootstrap. Now the Wald, F-Test and restricted bootstrap tests share the same power as we see in Figure 4. But it still holds that the unrestricted bootstrap exhibits higher power than the other tests.

The conclusion to our power investigation is that, generally, the unrestricted bootstrap test performs better than the parametric tests and the restricted bootstrap tests in all samples, and that its superiority to the other tests is much more obvious with small samples.

Figure 3 Size-Power Curves for Restricted vs. Unrestricted Bootstrap Tests

Figure 3a 25 Observations

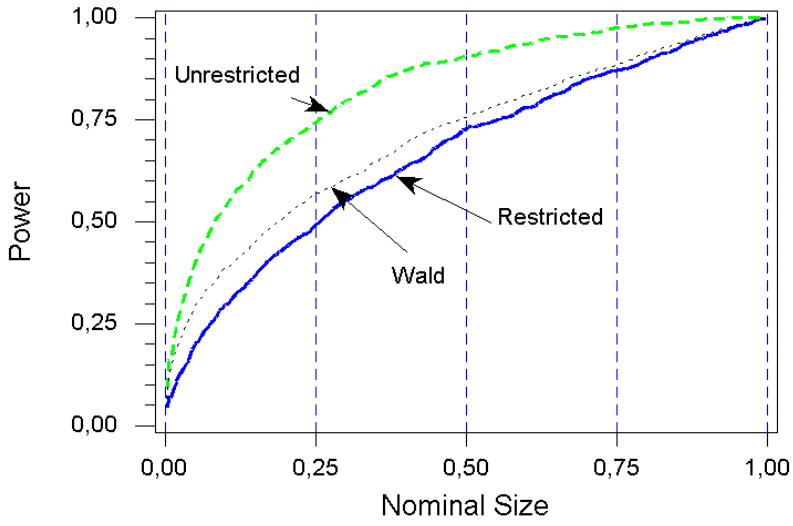


Figure 3b 50 Observations

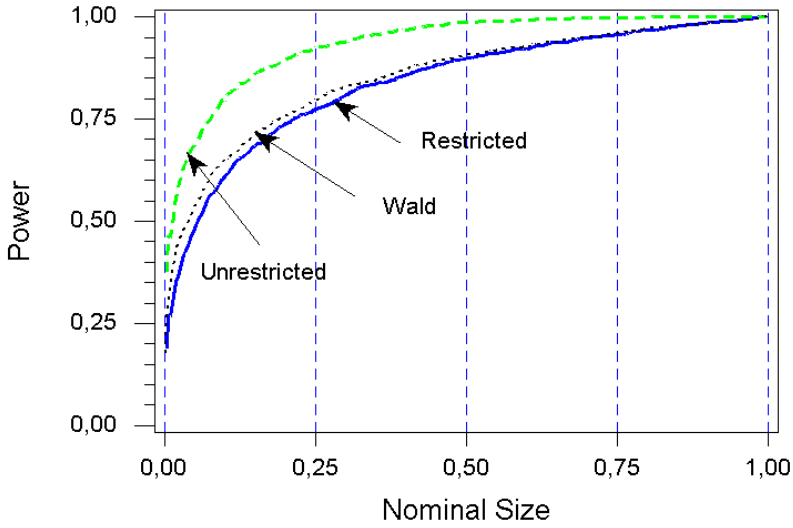


Figure 3c 100 Observations

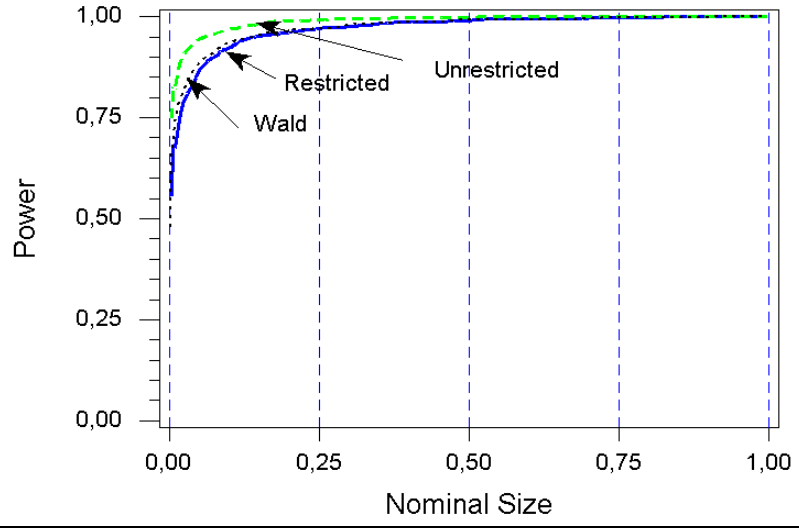


Figure 4: Size-Power Curves on a correct size-adjusted basis

Figure 4a 25 Observations

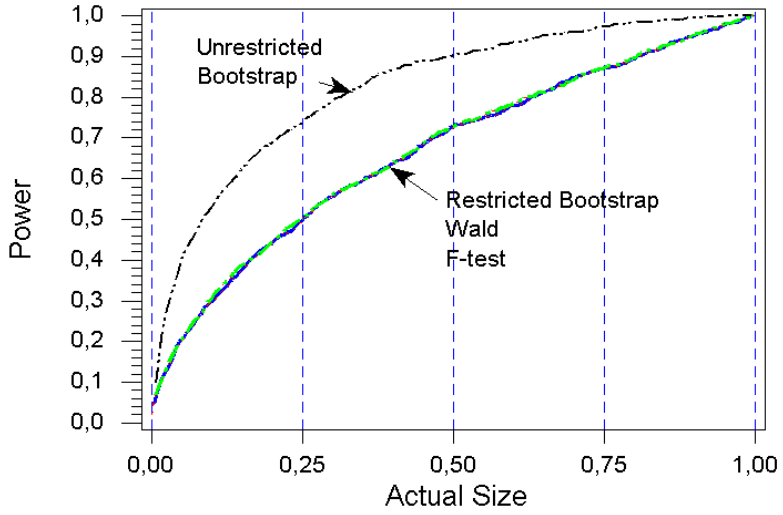


Figure 4b 50 Observations

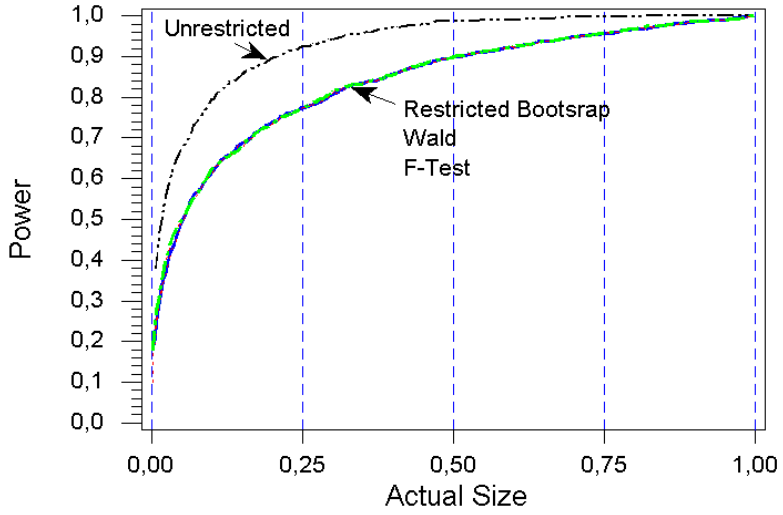
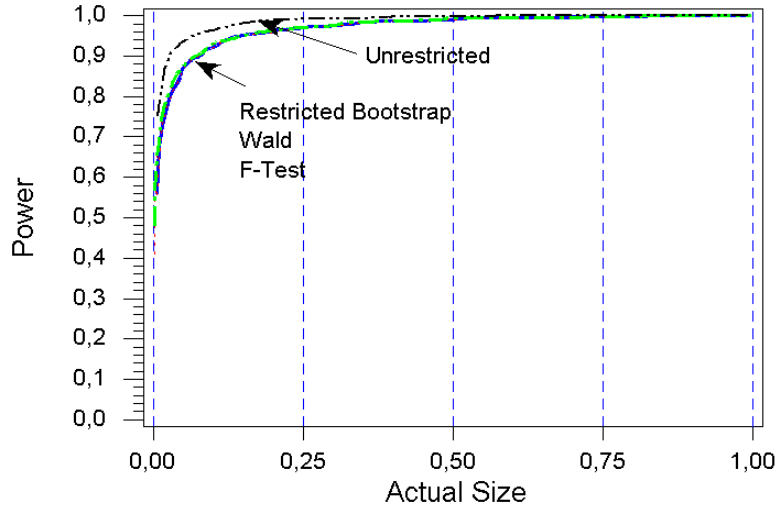


Figure 4c 100 Observations



6. Conclusions and Brief Summary

Let us summarize the results of our investigation. The purpose of this study has been to study the BG- autocorrelation test in a single equation dynamic model. We tested for autocorrelation of the first order, testing the hypothesis: $H_0 : \psi_1 = 0$, by using the bootstrap technique in two ways:

I) The restricted bootstrap test was used, in which we approximated the distribution of the test statistic, generating more robust critical values for our test statistic.

II) On the other hand, by the unrestricted bootstrap test, we approximated the distribution of the parameter (coefficient) ψ_1.

In both cases it does not matter whether or not we know the nature of the theoretical distribution of the parameter estimator or the theoretical distribution of the test statistic. What matters is that the bootstrap technique well approximates those distributions.

The conclusion to our investigation is that both bootstrap tests have an actual size that lies close to the nominal size and, given that the unrestricted model has the greatest power, it makes sense to choose the bootstrap from the unrestricted model ahead of the other tests, especially in small samples.

The difference between the two methods is that in the restricted (I) case we estimate and test indirectly the parameter (coefficient) ψ_1, through a test statistic, while in the unrestricted (II) case we estimate and test directly the parameter (coefficient) ψ_1. This second method has an advantage over the restricted (I) case, considering the power of the test in small samples. In large samples, however, the tests show a tendency to have the same power.
Moreover, while the restricted bootstrap test method can be easily generalized and is valid for a very general hypothesis about the serial correlation in the errors, not only in a single equation but also in systems of equations in the way that we have described. The unrestricted bootstrap, on the other hand, is not as clear as the restricted bootstrap test method, in the sense of generalizing the method for being valid for a very general hypothesis about the serial correlation in the errors and so we need to make further investigations. Even so, this cannot

diminish the fact that the unrestricted bootstrap test method is a powerful tool for testing the presence of autocorrelation of the first order in dynamic models.

In this paper we have studied the estimated size and power of the parametric and bootstrap tests by bootstrapping data under the null hypothesis (restricted bootstrap test) and under the alternative (unrestricted bootstrap test). Regarding the size of the test, we used Monte Carlo methods to investigate the properties of the tests using 10,000 replications per model, and we used the P-value plots to investigate the size of the tests. We found that the bootstrap tests perform better in small samples, while the restricted bootstrap is just a little better than the unrestricted bootstrap method in large samples.

When we consider the power results by studying the Size-Power Curves, even by studying the Size-Power Curves on a correct size-adjusted basis, the unrestricted model has the greater power. Finally, all tests share the same power, or the difference is very small for large samples.

References

Breusch, T. S. (1978). "Testing for Autocorrelation in Dynamic Linear Models," *Australian Economic Papers*, 17, 334-55.

Davidson, R. and J. G. MacKinnon (1996). "The Size Distortion of Bootstrap Tests," Working Paper, Department of Economics, Queen's University, Canada.

Davidson, R., and J. G. MacKinnon (1998). "*Graphical Methods for Investigating the Size and Power of Test Statistics*," The Manchester School, **66**, 1-26.

Durbin, J. and G. S. Watson (1950). "Testing for Serial Correlation in Least Squares Regression," *Biometrika* 37, 409-428

Edgerton, D. L. and G. Shukur (1999). "Testing Autocorrelation in a System Perspective," *Econometric Reviews*, 18(4), 343-386.

Freedman, D.A. (1984). "On Bootstrapping Two-Stage Least Squares Estimates in Stationary Linear Models," *Ann. Statist.* 12, 827-842.

Godfrey, L. G. (1978): "Testing for Higher Order Serial Correlation in Regression Equations When the Regressors Include Lagged Dependent Variables," *Econometrica,* 46, 1303-1310.

Horowitz, J. L. (1994. "Bootstrap-Based Critical Values for the Information Matrix Test," *Journal of Econometrics,* 61, 395-411.

Jeong, J. and S. Chung (2001): "Bootstrap Test for Autocorrelation" *Computational Statistics and Data Analysis,* 38(1) 49-69.

MacKinnon, J. G. (1992). "Model Specification Tests and Artificial Regressions," *Journal of Economic Literature,* 30, 102-146.

Maddala, G. S. (1992). *Introduction to Econometrics,* Second Edition, New York, Maxwell Macmillan.

Mantalos, P. and G. Shukur (1998). "Size and Power of the Error Correction Model (ECM) Cointegration Test - A Bootstrap Approach," *Oxford Bulletin of Economics and Statistics*, 60, 249-255.

Morey, M. J. and S. Wang (1985). "Bootstrapping the Durbin-Watson Statistic," *Proceedings of Business and Economic Statistics,* American Statistical Association, 3,549-553.

Rao, C. R. (1973): *Linear Statistical Inference and Its Applications*, Second edition. New York: Wiley.

Shukur, G. and P. Mantalos (1997a). "Size and Power of the RESET Test as Applied to Systems of Equations: A Bootstrap Approach," Working Paper 1997:3. Department of Statistics, University of Lund, Sweden.