Article:

Godfrey, L.G. (2007) Alternative approaches to implementing Lagrange multiplier tests for serial correlation in dynamic regression models. Computation Statistics & Data Analysis. pp. 3282-3295.

https://doi.org/10.1016/j.csda.2006.05.020

Published paper

Godfrey, L.G. (2007) *Alternative approaches to implementing Lagrange multiplier tests for serial correlation in dynamic regression models.* Computational Statistics & Data Analysis, 51 (7). pp. 3282-3295. ISSN 0167 9473

June 21, 2007

Alternative approaches to implementing Lagrange multiplier tests for serial correlation in dynamic regression models.

Abstract

An approximate F-form of the Lagrange multiplier test for serial correlation in dynamic regression models is compared with three bootstrap tests. In one bootstrap procedure, residuals from restricted estimation under the null hypothesis are resampled. The other two bootstrap tests use residuals from unrestricted estimation under an alternative hypothesis. A fixed autocorrelation alternative is assumed in one of the two unrestricted bootstrap tests and the other is based upon a Pitman-type sequence of local alternatives. Monte Carlo experiments are used to estimate rejection probabilities under the null hypothesis and in the presence of serial correlation.

Keywords: Bootstrap; Serial correlation; Lagrange multiplier test

L. G. Godfrey

Department of Economics, University of York, Heslington, York YO10 5DD, UK.
Tel: +44-1904-433754
Fax: +44-1904-433759
E-mail: lgg1@york.ac.uk

1 Introduction

The importance of testing for serial correlation in the error terms of a linear regression model has been recognized for many years. In general, the presence of serial correlation invalidates the classical formula for the variance-covariance matrix of ordinary least squares (OLS) estimators and leads to the inconsistency of these estimators if the regressors include lagged values of the dependent variable. The best-known and most widely reported test for serial correlation is the procedure proposed in Durbin and Watson (1950, 1951). However, as pointed out in many texts, e.g., Davidson and MacKinnon (2004, pp. 280-1) and Maddala (2001, p. 230), the Durbin-Watson (hereafter DW) test is subject to several limitations. In particular, the DW test uses a first-order autoregression as the alternative model and cannot be used when the regressors include lagged values of the dependent variable. While the emphasis on a simple first-order alternative may be justified when estimation is based upon annual data, as was probably often the case when the DW test was derived, there is the risk that power will be low when seasonal autocorrelation is present; see Godfrey and Tremayne (1988). Many time series regressions are now estimated using quarterly or monthly data so that this potential weakness is important. The second limitation is also important because, e.g., autoregressive distributed lag models are popular in applied work.

It is now standard practice to use the Lagrange Multiplier (LM) tests of Breusch (1978) and Godfrey (1978) when it is desired to allow for higher-order alternatives and the presence of lagged dependent variables in the regressors. The LM tests use the results of restricted estimation, i.e., estimation under the null hypothesis, but do not require the unrestricted estimation of the alternative model. Thus they have the convenient property of being based upon OLS results. However, the LM tests suffer from the drawback that they are only asymptotically valid and the asymptotic χ^2 critical values have sometimes been found to give inadequate control of finite sample significance levels; see, e.g., Kiviet (1986). Fortunately there is also evidence that improved control can be gained by a very simple modification of the original LM statistics.

As is widely known, LM tests for serial correlation can be interpreted as tests for omitted variables. The equation under test serves as the null model and lagged values of the residuals from its estimation are the additional variables that are to be tested for joint significance in an artificial alternative regression; see Breusch and Godfrey (1981) and Davidson and MacKinnon (2004, Section 7.7). The standard F test is not only a natural choice for this purpose, but it has also been found in Kiviet (1986) to lead to improved finite sample performance relative to asymptotic χ^2 variants. However, there have been enormous advances in the power of computers since the work of Kiviet (1986) and it is now feasible to consider the application of resampling methods to improve upon the approximation given by asymptotically valid critical values.

Resampling approaches to controlling the finite sample rejection rates of tests have been discussed for several different types of checks of the specification of regression models. Important examples of such checks are the information test of White (1982) and the J-test of Davidson and MacKinnon (1981). For both of these tests, evidence has been produced to show that bootstrap critical values are much more reliable than those obtained from asymptotic theory; see, e.g., Horowitz (1994) on the bootstrap information test and Godfrey (1998) for the bootstrap J-test. It is, therefore, not surprising that the bootstrap has been applied to the serial correlation tests proposed in Breusch (1978), Durbin (1970) and Godfrey (1978) for use in dynamic regression models.

The first such application seems to be Rayner (1993). Rayner uses a form of unrestricted (alternative hypothesis) residual resampling. The choice of the estimation method used to obtain unrestricted residuals for bootstrap schemes requires care when the regressors include lagged dependent variables. In the standard theory that underpins the bootstrap test, as discussed in Beran (1988), it is assumed that estimators of nuisance parameters are consistent. However, under the serial correlation alternative, OLS estimators are usually inconsistent when the regressors include lagged values of the dependent variable. Consequently OLS residuals are generally inappropriate for resampling under the alternative. Rayner uses an instrumental variable (IV) technique that is consistent whether or not there is serial correlation; see the Appendix of Rayner (1993) for details. After conducting Monte Carlo experiments, Rayner finds evidence that asymptotic critical values are inadequate and bootstrap-based tests are well behaved under the null hypothesis.

An approach to obtaining an unrestricted bootstrap scheme that does not require IV estimation is proposed in Mantalos (2003); also see Mantalos (2005). The unrestricted residuals used in Mantalos (2003) are not derived from the actual alternative model that has been specified but instead from the artificial regression mentioned above that is used to calculate the test statistic. The latter model is locally equivalent to the former in a sense that is explained below. Encouraging results on this simpler approach are reported in Mantalos (2003).

The alternative hypothesis used in Mantalos (2003) and Rayner (1993) is the classical first-order autoregressive error model. This alternative is also adopted with a dynamic regression equation in an experiment reported by MacKinnon; see MacKinnon (2002, pp. 623-624). In MacKinnon's work, restricted residuals from OLS estimation of the null model are used for the bootstrap. It is found that asymptotic critical values can be quite unreliable while bootstrap critical values produce good performance.

Overall, given the existing results, it appears that, whether restricted or unrestricted residuals are used, bootstrap versions of serial correlation tests are well-behaved under the null hypothesis. Moreover, power comparisons seem to give some support to the use of unrestricted residuals in preference to restricted residuals; see Mantalos (2003). However, there are issues that indicate the need for additional research and the following are addressed in this paper.

First, the existing studies have concentrated on experimental designs in which, under the alternative, the errors are normally distributed and follow a first-order autoregression. This restrictive error model is, however, the one used in Durbin and Watson (1950, 1951). There seems little justification for adopting it when examining bootstrap versions of LM tests that were developed for use in more general circumstances. One purpose of this paper is, therefore, to provide evidence for higher-order autocorrelation models with nonnormal disturbances.

Second, the comparison and implementation of restricted and unrestricted bootstrap tests also merit further consideration. It is argued below that when using checks for misspecification, such as serial correlation tests, it must be acknowledged that it is unlikely that there will be precise information about the alternative hypothesis and moreover that a given LM statistic could be derived from any member of a family of locally equivalent alternatives. These points have relevance when an unrestricted bootstrap is used. In the unrestricted bootstrap algorithm used in Jeong and Chung (2001) and in Rayner (1993), an autoregression of the same order as the selected alternative is fitted to the OLS residuals to derive the terms to be used in resampling. It is possible that the order of the selected alternative will be incorrect. Moreover, even if the order has been selected correctly, the use of an autoregression may lead to misleading inferences if the true error process is moving average. Consequently it is important to investigate the robustness of autoregression-based unrestricted bootstrap methods to misspecification of the alternative model.

The plan of the paper is as follows. The models and tests are discussed in Section 2. The design of simulation experiments is described in Section 3. The results from the experiments are summarized in Section 4. Finally, Section 5 contains some concluding remarks.

2 Regression models and serial correlation tests

Suppose that data are generated by the stable dynamic linear model

$$y_t = Y_t'\alpha + X_t'\beta + u_t = W_t'\gamma + u_t, \tag{1}$$

where: $Y_t^{'} = (y_{t-1}, ..., y_{t-L})$ with $L \geq 1$; $\alpha' = (\alpha_1, ..., \alpha_L)$ has elements such that the roots of

$$z^L - \alpha_1 z^{L-1} - ... - \alpha_L = 0,$$

are all strictly inside the unit circle; X_t is an M-vector of regressors that are strictly exogenous; $W_t' = (Y_t', X_t')$; and $\gamma' = (\alpha', \beta')$. Let $K = L + M$ denote the number of regression coefficients in (1) and N denote the number of observations available for estimation. The errors u_t are assumed to be covariance stationary with zero mean and variance σ^2, with common cdf F_u. It is not assumed that the errors are normally distributed. As in Breusch (1978), Durbin (1970) and Godfrey (1978), it is convenient to assume that

$$plim N^{-1} \sum_{t=1}^{N} W_t W_t'$$

is a finite positive-definite matrix. It is, however, possible to allow the exogenous regressors to be nonstationary without compromising the asymptotic validity of serial correlation tests; see Wooldridge (1999).

The null hypothesis to be tested is that the errors u_t are serially uncorrelated. All tests are constructed using the results of OLS estimation of (1). Let $\hat{\gamma}' = (\hat{\alpha}', \hat{\beta}^{'})$ denote the OLS coefficient estimator for (1) and the terms $\hat{u}_t = y_t - W_t'\hat{\gamma}$ be the corresponding residuals, $t = 1, ..., N$. Under the assumptions and the null hypothesis, the OLS estimators are consistent with $(\hat{\gamma} - \gamma)$ being $O_p(N^{-1/2})$. Whether the alternative is a Gth-order autoregression, denoted by AR(G) and written as

$$u_t = \sum_{j=1}^{G} \phi_j u_{t-j} + \epsilon_t, \epsilon_t iid(0, \sigma_\epsilon^2), \tag{2}$$

or a Gth-order moving average, denoted by MA(G) and written as

$$u_t = \sum_{j=1}^{G} \theta_j \epsilon_{t-j} + \epsilon_t, \epsilon_t iid(0, \sigma_\epsilon^2), \tag{3}$$

a suitable LM test can be computed as test of $\lambda = (\lambda_1, ..., \lambda_G)' = 0$ in the augmented model

$$y_t = Y_t'\alpha + X_t'\beta + \hat{U}_t'\lambda + u_t = W_t'\gamma + \hat{U}_t'\lambda + u_t, \tag{4}$$

in which $\hat{U}_t' = (\hat{u}_{t-1}, ..., \hat{u}_{t-G})$ and $\hat{u}_{t-g}$ is set equal to zero for $t \leq g$. Let the OLS estimators for (4) be $\tilde{\gamma} = (\tilde{\alpha}', \tilde{\beta}')'$ and $\tilde{\lambda}$. The LM test is then a check of the joint significance of the elements of $\tilde{\lambda}$. A large sample test based upon the χ^2_G distribution is used in Breusch (1978) and Godfrey (1978), but the results in Kiviet (1986) indicate that the (asymptotically valid) standard F-test of (1) versus (4) is better behaved in finite samples. The latter test is denoted by LM$_F$.

Since the limit null distribution of LM$_F$ is χ^2_G/G, the F-statistic of Kiviet (1986) is asymptotically pivotal, i.e., its asymptotic distribution is independent of the nuisance parameters, which are taken to include the error distribution function F_u. The results of Beran (1988), therefore, indicate that bootstrap tests may yield more accurate inferences than the asymptotic checks. A general bootstrap scheme can be written as

$$y_t^* = \sum_{j=1}^{L} y_{t-j}^* \ddot{\alpha}_j + X_t'\ddot{\beta} + u_t^*, t = 1, ..., N, \tag{5}$$

in which: (i) presample values of y^* are set equal to those of y; (ii) under the null hypothesis, $\ddot{\alpha} = (\ddot{\alpha}_1, ..., \ddot{\alpha}_L)'$ and $\ddot{\beta}$ are both consistent; and (iii) under the null hypothesis, the distribution function of the bootstrap errors u_t^* converges in an appropriate (Mallows) metric to that of the true errors u_t. The choice of consistent estimator for the regression coefficients and the choice of scheme used to obtain u_t^* may both have small sample effects that cannot be neglected. The approaches adopted in the literature can be summarized as follows.

First, perhaps the most natural way in which to mimic the true data process under the null hypothesis is to use the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ from (1) for $\ddot{\alpha}$ and $\ddot{\beta}$, respectively, with the bootstrap errors u_t^* being obtained by simple random sampling with replacement from the empirical distribution function

$$\hat{F}_u^r : \text{probability } \frac{1}{N} \text{ on } \hat{u}_t, t = 1, ..., N. \tag{6}$$

This combination gives the restricted (null hypothesis) bootstrap model

$$y_t^* = \sum_{j=1}^{L} y_{t-j}^* \hat{\alpha}_j + X_t'\hat{\beta} + u_t^*, t = 1, ..., N, \tag{7}$$

with u_t^* being derived using (6). If (1) does not contain an intercept term, the sample mean of the OLS residuals should be subtracted from each $\hat{u}_t$ before it is used in (6). Mean-adjustment is also required if the OLS residuals $\hat{u}_t$ are modified by being divided by the square root of $(1 - h_{tt})$, where h_{tt} is the leverage value.

In contrast to the OLS-based approach of (7), IV estimators are used in Rayner (1993) for $\ddot{\alpha}$ and $\ddot{\beta}$ with the instruments consisting of current and lagged exogenous variables. With this choice of instruments, the estimators are also consistent under a fixed alternative hypothesis, i.e., in the unrestricted model. Under a fixed AR(G) (resp. MA(G)) alternative, at least one of the coefficients ϕ_j of (2) (resp. θ_j of (3)) is a nonzero constant. In general, the statistic LM$_F$ is O$_p(N)$ under such an alternative so that the asymptotic rejection probability tends to unity for any finite critical value.

Let the IV parameter estimators and residuals be denoted by $\check{\alpha}$, $\check{\beta}$ and $\check{u}_t$, respectively. In order to derive an unrestricted bootstrap scheme, it remains to specify how to obtain u_t^*. An AR(1) alternative is assumed in Rayner (1993) and a generalization for the AR(G) case involves applying OLS to

$$\check{u}_t = \phi_1 \check{u}_{t-1} + ... + \phi_G \check{u}_{t-G} + e_t. \tag{8}$$

Let the residuals derived from OLS estimation of (8) be denoted by $\check{e}_t$. The unrestricted (alternative hypothesis) bootstrap model is

$$y_t^* = \sum_{j=1}^{L} y_{t-j}^* \check{\alpha}_j + X_t' \check{\beta} + u_t^*,$$

with u_t^* being obtained by simple random sampling with replacement from the empirical distribution function

$$\check{F}_u^u : \text{probability } \frac{1}{N} \text{ on } \check{e}_t^c, t = 1, ..., N, \tag{9}$$

where $\check{e}_t^c$ is the centred (demeaned) version of $\check{e}_t$. The bootstrap test LM_F^* is then calculated using these artificial observations y_t^* and the bootstrap counterparts of (1) and (4).

Since this unrestricted bootstrap requires both IV and OLS estimation of (1), and the OLS fitting of (8), it has a higher computational cost than the restricted bootstrap. The asymptotic theory of Beran (1988) does not predict that the unrestricted method will enjoy any advantage in terms of approximating the required significance level. (Indeed, in a study of parametric bootstrap tests applied to asymptotically normal pivots, Lee and Young argue that nuisance parameters should be replaced by their restricted estimates; see Lee and Young, 2005.) As noted in MacKinnon (2002, p. 621), some researchers have argued that unrestricted bootstrap schemes will produce tests with better power than those from restricted schemes, but the available Monte Carlo evidence is mixed. Also there are difficulties with applying such arguments to tests for serial correlation (and more generally to tests for misspecification).

In the unrestricted bootstrap model approach of Rayner (1993), the parameters of the conditional bootstrap law are estimators from observed data that are consistent in the specified unconditional fixed alternative model. This feature might be thought to yield residuals that give a better approximation to the distribution of the error terms under the alternative. However, it is well-known that there is more than one alternative that leads to LM_F; see, e.g., Godfrey (1988, Section 4.4.1). Consequently the use of the fitted autoregression (8) may be invalid under fixed alternatives. If the u_t were generated by a MA(G) process, the residuals $\check{e}_t^c$ used in (9) would be inappropriate; see Schwert (1987) for comments on the dangers of relying on pure autoregressions. Ramsey's criticism of the use of specific alternative models seems pertinent in the context of serial correlation tests; see Ramsey (1983, p. 243-244). Thus there must be doubts about the general usefulness of the unrestricted bootstrap of Rayner (1993).

These doubts expressed about the use of an unrestricted bootstrap with LM_F, or with any other test for misspecification, should not be taken as applying to the procedures studied in van Giersbergen and Kiviet (2002). The alternative model in van Giersbergen and Kiviet (2002) is a regression model such as (1) and the null model is derived by imposing linear restrictions on its coefficients. In this framework, there is no need to worry about a family of locally equivalent alternatives and the paucity of information to guide the correct choice from this family.

The second type of unrestricted bootstrap to be considered is based upon a scheme described in

Mantalos (2003). The alternative used in Mantalos (2003) is the AR(1) model. A generalization that allows for the Gth-order alternative consists of the following steps.

1. Estimate (1) by OLS to obtain $\hat{\alpha}$, $\hat{\beta}$ and the residuals $\hat{u}_t$.
2. Estimate (4) by OLS to obtain $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\lambda}$ and the residuals $\tilde{u}_t$.
3. Draw $e_1^*, ..., e_N^*$ by simple random sampling with replacement from the empirical distribution function

$$\tilde{F}_u^u : \text{probability } \frac{1}{N} \text{ on } \tilde{u}_t^c, t = 1, ..., N, \tag{10}$$

where $\tilde{u}_t^c$ is the centred (demeaned) version of $\tilde{u}_t$: asymptotically negligible modifications of the latter residuals (as described, e.g., by MacKinnon, 2002, p. 620) are used in Mantalos (2003).

4. Given $e_1^*, ..., e_N^*$ from step 3, generate the bootstrap errors u_t^* recursively using

$$u_t^* = \sum_{j=1}^{G} \tilde{\lambda}_j u_{t-j}^* + e_t^*, \tag{11}$$

with required starting values set equal to zero.

5. Generate bootstrap data using

$$y_t^* = \sum_{j=1}^{L} y_{t-j}^* \tilde{\alpha}_j + X_t' \tilde{\beta} + u_t^*, t = 1, ..., N, \tag{12}$$

in which the bootstrap errors are given by (11).

6. The bootstrap value of the Breusch-Godfrey LM$_F$ statistic, denoted by LM$_F^*$, is then obtained by testing $H_0^* : \lambda = \tilde{\lambda}$, not $H_0 : \lambda = 0$, in the bootstrap counterpart of (4); see van Giersbergen and Kiviet (2002, Section 1.2). The OLS estimators of the coefficients of the bootstrap version of the artificial alternative (4) are denoted by $\tilde{\gamma}^*$ and $\tilde{\lambda}^*$.

In the version of an unrestricted bootstrap given by steps 1 to 6, there is no need to employ IV, as well as OLS, estimation. This saving reflects the fact that (4) is being used as an approximation to the specified alternative. The approximation is asymptotically valid under local alternatives in which parameters that are under test are $O(N^{-1/2})$, rather than $O(1)$; see Godfrey (1981). Thus the coefficients that determine the pattern of error autocorrelation are given by a Pitman-type drift, rather than being fixed constants. Under an artificial sequence of alternatives that are drifting towards the null model at the specified rate in the unconditional (real world) law, the OLS estimators for (4) are consistent and, in particular, $\tilde{\lambda}$ tends to a null vector so that (11) represents a local alternative in the conditional (bootstrap) world. However, the local asymptotic theory of Pitman drifts may provide a poor approximation to actual behaviour in finite samples drawn from the bootstrap population when observed serial correlation is not weak; see Eastwood and Godfrey (1992, Section 4.2) for a pilot Monte Carlo study on such approximations for the actual population.

In order to gain some insight into the possible effects of poor approximations in the scheme of Mantalos (2003), consider the consequences of interpreting $\tilde{\lambda}$ in the bootstrap world as being a fixed vector of constants, rather than being $O(N^{-1/2})$. As is clear from steps 4 and 5, the bootstrap data are generated using a model that is characterized by population coefficient vectors of $\tilde{\gamma} = (\tilde{\alpha}', \tilde{\beta}')'$ in the regression mean function and $\tilde{\lambda}$ in the AR(G) error model. The combination of serial correlation generated by (11) and lagged dependent variables y_{t-j}^* included as regressors in (12) implies that, under fixed autocorrelation alternatives, the OLS estimators of the bootstrap counterpart of (4) are inconsistent with.

$$plim^* \tilde{\lambda}^* \neq \tilde{\lambda},$$

in which $plim^*$ denotes a bootstrap probability limit taken as $N \longrightarrow \infty$ under (11) and (12). As a result of this inconsistency, the bootstrap test statistic LM_F^* is $O_p^*(N)$ when $H_0^* : \lambda = \tilde{\lambda}$ is true, where $O_p^*(\cdot)$ denotes a stochastic order of magnitude in the bootstrap world. In contrast, under $H_0 : \lambda = 0$, LM_F is $O_p(1)$ with its limit null distribution being $\chi^2(G)/G$. Thus, if fixed autocorrelation model parameters were assigned to the bootstrap world, the conditional law probability that LM_F^* exceeds LM_F would tend to unity as $N \longrightarrow \infty$, along almost all sample sequences. Since $H_0 : \lambda = 0$ is to be rejected when the estimated p-value of LM_F is small, the unrestricted procedure based on steps 1-6 may lead to the true null hypothesis being rejected less frequently than desired in finite samples, if asymptotic local theory fails to provide a good approximation in the bootstrap world.

3 Monte Carlo design

The Monte Carlo data generation process corresponding to (1) is obtained using $L = M = 2$; so that $K = 4$. It is written as

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta_1 + \beta_2 x_t + u_t, t = 1, ..., N, \tag{13}$$

in which N is either 40 or 80. Under the null, the u_t are iid$(0, \sigma^2)$. This process has been used in Dezhbakhsh (1990), Dezhbakhsh and Thursby (1995) and Godfrey and Tremayne (2005); parameter values in (13) are specified as in these earlier papers. The values of (α_1, α_2) are (0.5, 0.3), (0.7, -0.2), (1.0, -0.2), (1.3, -0.5), (0.9, -0.3) and (0.6, 0.2), which all satisfy the conditions for dynamic stability. The value of (β_1, β_2) is (1, 1) in all cases. The values of σ^2 are 1, 10 and 100. The OLS estimates of the parameters of (13) are denoted by $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1$ and $\hat{\beta}_2$.

The null hypothesis of serially uncorrelated errors is tested with $G = 4$; so that the test model corresponding to (4) is

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta_1 + \beta_2 x_t + \sum_{j=1}^{4} \lambda_j \hat{u}_{t-j} + error, \tag{14}$$

in which the terms $\hat{u}_{t-j}$ are the lagged residuals from the OLS estimation of (13). The OLS estimate of (α_1, α_2) from (14) is denoted by $(\tilde{\alpha}_1, \tilde{\alpha}_2)$. A test of (13) against (14) corresponds to the kind of serial correlation check that might be used when working with quarterly data.

Genuine data are used for values of the exogenous regressor x_t in Rayner (1993). However, this approach implies that this regressor is fixed in repeated sampling over the Monte Carlo replications. Given that the aim is to obtain Monte Carlo evidence relevant to non-experimental regression data analysis, this implicit conditioning seems undesirable. As in several other studies, x_t is generated by a stable AR(1) process, i.e.,

$$x_t = \rho_x x_{t-1} + \zeta_t, |\rho_x| < 1, \zeta_t \ NID(0, \sigma_\zeta^2). \tag{15}$$

In experiments that are not reported here, it is found that using a simple AR(4) model,

$$x_t = \rho_x x_{t-4} + \zeta_t, |\rho_x| < 1, \zeta_t \ NID(0, \sigma_\zeta^2), \tag{16}$$

does not lead to important differences in results. The results discussed below are obtained using $\rho_x = 0.7$ and σ_ζ^2 selected so that $Var(x_t) = 1$. In order to obtain IV estimators for (13) that are consistent under a serial correlation alternative, x_t, x_{t-1}, x_{t-2} and an intercept term are used as instruments. The corresponding IV estimate of (α_1, α_2) from (13) is denoted by $(\breve{\alpha}_1, \breve{\alpha}_2)$.

The error terms u_t of (13) are generated by special cases of the mixed autoregressive-moving average ARMA(5, 5) model

$$u_t = \sum_{j=1}^{5} \phi_j u_{t-j} + \epsilon_t + \sum_{j=1}^{5} \theta_j \epsilon_{t-j}, \tag{17}$$

in which the ϵ_t are independently and identically distributed (iid) with zero mean and variance σ_ϵ^2. All pre-sample values for (17) are set equal to zero. Similarly starting values for y and x are set equal to their respective unconditional means. The effects of this standard computational device are reduced by generating $N + 50$ observations and then using the last N of them.

Model (17) is sufficiently general to provide evidence about several aspects of the performance of the asymptotic and bootstrap tests derived from (13) and (14). By appropriate choices of the coefficients of (17), rejection rates can be estimated: under the null hypothesis; under an AR(4) scheme, as is used in the unrestricted bootstrap tests; and under serial correlation models that are not AR(4).

The iid error term ϵ_t of (17) is drawn from three distributions. Following the previous studies of restricted and unrestricted bootstrap serial correlation tests, the Normal distribution is used. The other two choices give symmetric and asymmetric nonnormal distributions. The former involves drawing iid terms from a $t(5)$ distribution and then transforming them to have the required population mean and variance. For the latter, the $\chi^2(8)$ distribution is the source of the drawings that are transformed.

Tests are implemented in p-value form. The p-value for the procedure of Kiviet (1986) is

$$FPV = \Pr(\mathrm{F}(4, N-8) \geq \mathrm{LM}_F^{obs}),$$

for $N = 40, 80$, in which LM_F^{obs} is the observed value of LM_F. The estimated p-values for the restricted, Mantalos-type unrestricted and Rayner-type unrestricted bootstrap tests are denoted by RES, MUR and RUR, respectively. These bootstrap p-values are calculated using 1000 bootstrap samples. Rejection rates are obtained by comparing calculated p-values with nominal significance levels of 5% and 10%. Estimates of rejection probabilities are based upon a maximum of 25000 replications. The asymptotic test of Kiviet (1986) is available in every replication. However, some replications are not suitable for computing bootstrap tests. The problem is that estimates of (α_1, α_2) may fail to satisfy the conditions for dynamic stability and so cannot be used to define covariance stationary bootstrap data generation processes. The Monte Carlo results concerning this problem and the finite sample rejection rates of the various tests under null and alternative hypotheses are discussed in the next section.

4 Monte Carlo results

In the discussion of the Monte Carlo results, it is useful to have a code for combinations of (α_1, α_2) and the error distribution. The code system is given in Table 1. For example, a case with code 2B has $(\alpha_1, \alpha_2) =$ (0.7, -0.2) and errors derived from the $t(5)$ distribution.

Table 1

Codes for combination of (α_1, α_2) and error distribution	
1	$(\alpha_1, \alpha_2) = (0.5, 0.3)$
2	$(\alpha_1, \alpha_2) = (0.7, -0.2)$
3	$(\alpha_1, \alpha_2) = (1.0, -0.2)$
4	$(\alpha_1, \alpha_2) = (1.3, -0.5)$
5	$(\alpha_1, \alpha_2) = (0.9, -0.3)$
6	$(\alpha_1, \alpha_2) = (0.6, 0.2)$
A	error distribution from Normal
B	error distribution from $t(5)$
C	error distribution from $\chi^2(8)$

Before examining estimates of significance levels, consider first the results that shed light on how frequently bootstrap tests are applicable. The proportion of replications in which the estimates of (α_1, α_2) required for defining the bootstrap population are consistent with stationarity is an obvious index of applicability. Such proportions, measured in percentage terms, are referred to as applicability ratios. Perusal of the results indicates that applicability ratios are not very sensitive to the choice of error distribution from A, B and C of Table 1. As a representative sample, results on applicability for cases with Normal errors are reported in Table 2. Table 2 has 18 groups of results, each corresponding to a combination of bootstrap test (3 types), N (2 values) and σ^2 (3 values). There are 6 applicability ratios in each group, corresponding to the combinations of (α_1, α_2) in Table 1.

Table 2
Applicability ratios (percentages) for Normal errors

$N = 40$ with	$\sigma^2 = 1$	$\sigma^2 = 10$	$\sigma^2 = 100$
RES test	100.0,100.0,100.0, 100.0,100.0,100.0	99.9,100.0,100.0, 100.0,100.0,99.9	99.8,100.0,100.0, 100.0,100.0,99.9
RUR test	90.9,96.6,92.1, 86.9,95.0,92.4	49.7,58.7,53.7, 49.7,57.7,51.4	33.4,38.2,35.5, 36.3,38.0,33.6
MUR test	97.8,99.8,99.4, 99.7,99.8,98.4	72.3,79.7,79.8, 86.0,82.8,73.7	58.3,63.7,64.5, 75.3,68.5,58.1

$N = 80$ with	$\sigma^2 = 1$	$\sigma^2 = 10$	$\sigma^2 = 100$
RES test	100.0,100.0,100.0, 100.0,100.0,100.0	100.0,100.0,100.0, 100.0,100.0,100.0	100.0,100.0,100.0, 100.0,100.0,100.0
RUR test	98.7,99.6,99.1, 96.0,99.2,99.0	63.2,74.1,67.1, 62.0,71.3,65.4	36.5,42.3,39.0, 38.7,41.8,36.2
MUR test	99.9,100.0,100.0, 100.0,100.0,100.0	85.7,91.1,92.1, 95.8,92.8,87.6	65.2,67.6,72.1, 84.1,74.9,65.3

As can be seen from Table 2, there are important differences between the applicability ratios for the three different bootstrap tests. The restricted bootstrap check RES is almost always available whatever the combination of N and σ^2. The unrestricted bootstrap tests, however, fail to match this level of performance. Instead the value of σ^2 is associated with substantial effects. As σ^2 increases so does the relative frequency with which estimates of (α_1, α_2) imply dynamically unstable bootstrap data processes. The problems with the IV-based procedure RUR derived from Rayner (1993) are so marked that it is difficult to recommend it as a tool for general use in applied work. The applicability of the Mantalos-type test MUR is not so severely impaired by error variance increases, but the effects of such increases are not negligible.

The sensitivity of unrestricted bootstrap tests to variations in σ^2 can be discussed after rewriting (13) as

$$y_t = \Psi(B)s_t + \sigma\Psi(B)a_t, \tag{18}$$

in which: B is the backward shift operator, with $B^j y_t = y_{t-j}$; $\Psi(B) = 1 + \psi_1 B + \psi_2 B^2 + ... = (1 - \alpha_1 B - \alpha_2 B^2)^{-1}$; $s_t = \beta_1 + \beta_2 x_t$; and $a_t = \sigma^{-1} u_t$. In this representation of the data process, s_t and a_t are uncorrelated with $Var(s_t) = Var(a_t) = 1$ in all experiments. It follows that, as σ increases, the importance of the exogenous component $\Psi(B)s_t$ decreases relative to $\sigma\Psi(B)a_t$.

In the experiments, Rayner's version of the unrestricted bootstrap test uses x_{t-1} and x_{t-2} as instruments for y_{t-1} and y_{t-2}. From (18), these instruments are only correlated with the component $\Psi(B)s_{t-j}$ of $y_{t-j}, j = 1, 2$, Hence, as increases in σ make the exogenous components less and less important, a type of weak instruments problem is approached and it is not surprising that IV estimates are not close to the corresponding true parameter values.

A different explanation is required for the sensitivity of the Mantalos-type check because it does not use IV estimation. The unrestricted bootstrap test that is proposed by Mantalos uses OLS estimators of

(14) to define the bootstrap process. Under the null hypothesis, the estimators from (14) are inefficient relative to those for (13): the latter provide the parameter values for the restricted bootstrap. The degree of asymptotic variance inflation depends upon the extent to which y_{t-1} and y_{t-2} are "explained" in linear regressions by $u_{t-1}, u_{t-2}, u_{t-3}$ and u_{t-4}. Consequently (18) implies that the effects of asymptotic variance inflation increase as σ increases. These effects may be reflected in finite samples by the greater frequency with which $(\tilde{\alpha}_1, \tilde{\alpha}_2)$ from (14) imply nonstationary AR(2) regression models.

The fact that the test MUR is available more often than the test RUR is not sufficient to imply that the former is either well-behaved or superior to the latter. It is important to investigate the differences between actual and desired null rejection probabilities. When the Monte Carlo data process is such that all tests are usually available, the estimates for MUR are persistently below the desired values and the estimates for RUR suggest much closer agreement. The possibility of low rejection rates for MUR was discussed above and it appears that, in these experiments, AR error processes in the bootstrap world are not adequately approximated by the artificial alternative derived by adding lagged residuals to the original regression model. Table 3 contains results that illustrate these findings.

Table 3

Estimated null rejection probabilities, with $\sigma^2 = 1$ and nominal significance level of 5%

	$N = 40$				$N = 80$			
	24114 $\geqslant$ # replications $\geqslant$ 21707				24899 $\geqslant$ # replications $\geqslant$ 23994			
	FPV	RES	RUR	MUR	FPV	RES	RUR	MUR
Case 1A	4.6	4.2	4.3	2.3	5.3	5.1	5.1	2.9
Case 1B	4.5	4.2	4.2	2.3	4.8	4.9	4.9	2.7
Case 1C	4.5	4.3	4.3	2.3	4.8	4.8	4.8	2.6
Case 2A	4.5	4.6	4.6	2.0	4.9	5.0	5.0	2.7
Case 2B	4.2	4.6	4.5	1.9	4.5	4.9	4.9	2.7
Case 2C	4.2	4.4	4.3	1.9	4.6	5.0	4.9	2.6
Case 3A	5.1	4.8	4.7	2.8	4.9	4.8	4.7	3.2
Case 3B	4.4	4.3	4.1	2.5	4.6	4.8	4.7	3.3
Case 3C	4.6	4.4	4.3	2.5	4.8	4.8	4.8	3.2
Case 4A	5.0	4.3	4.3	3.3	5.1	4.9	4.9	4.4
Case 4B	4.7	4.3	4.2	3.4	4.7	4.7	4.8	4.1
Case 4C	4.9	4.3	4.3	3.4	4.9	4.9	4.8	4.1
Case 5A	4.7	4.7	4.6	2.2	4.7	4.7	4.8	2.9
Case 5B	4.1	4.3	4.2	2.2	4.7	5.0	5.0	3.2
Case 5C	4.3	4.6	4.3	2.0	4.4	4.7	4.6	2.8
Case 6A	5.3	4.9	4.8	2.6	4.9	4.8	4.8	2.7
Case 6B	4.6	4.4	4.2	2.4	4.8	5.0	4.9	2.8
Case 6C	4.8	4.5	4.5	2.3	4.8	5.0	4.9	2.6

Note: The case codes given in the first column are derived from the codes of Table 1.

The results in Table 3 are for the data processes defined by combining the cases of Table 1 with $\sigma^2 = 1$. The use of $\sigma^2 = 1$ implies that all tests are available with quite high frequency. Only replications in which all tests are available are used to obtain the results of Table 3. Consequently the number of replications varies with case and sample size. The smallest and largest numbers of such replications are shown for each sample size in Table 3. At worst, there are over 21000 replications, so that estimation should be sufficiently precise for practical purposes.

Rejection rates in Table 3 are derived with $\sigma^2 = 1$ so that all four tests can be compared. However, these results cannot be assumed to be representative of those for more general situations in which unrestricted bootstrap tests are not free of applicability problems. Attention is therefore also given to estimates for FPV and RES derived using all three values of σ^2. Table 4 contains estimates for RES and FPV for all values of σ^2, with $N = 40$. As indicated by Table 2, RES is almost always available and no estimate for this test in Table 4 is based upon fewer than 24948 replications (many are based upon the full set of 25000 replications). It is clear that RES performs well in the experiments. There is no indication of it being either persistently undersized or persistently oversized and fluctuations about the nominal size of 5% are small. Every estimate for RES is in the range 0.9×5% = 4.5% to 1.1×5% = 5.5%; so that all satisfy the stringent criterion of robustness given in Serlin (2000). The F-test approach of Kiviet (1986) does not seem to provide such good control of finite sample null rejection probabilities. Estimates for FPV fall on both sides of 5% when $\sigma^2 = 1$; but these estimates exhibit a greater degree of variability than the corresponding values for RES. For the two larger values of σ^2, the estimates for FPV provide evidence of a tendency for the test to be undersized, with the absolute value of the error in rejection probability increasing with σ^2. Comparison of the last two columns of results in Table 4 shows that the restricted bootstrap can be useful in correcting the test FPV, which relies upon the standard F-distribution for approximations.

Repeating the experiments of Table 4 with $N = 80$ does not lead to important changes. The same general patterns emerge and, for example, the average of estimates for FPV corresponding to those of the last column of Table 4 is 4.4%, compared with 4.3% when $N = 40$. The restricted bootstrap test has a slightly better performance when $N = 80$, with the largest difference between an estimate and the target value of 5% being 0.3%, compared with 0.4% when $N = 40$.

To sum up, consideration of the results for experiments in which the null hypothesis is imposed leads to the following conclusions. First, the use of critical values from an approximating F-distribution, as suggested in Kiviet (1986), can lead to underrejection relative to desired levels. Second, doubt is cast upon the general usefulness of the two unrestricted bootstrap tests either because of the possibility of being frequently inapplicable or because of excessively low rejection rates. Third, the restricted bootstrap leads to good control of finite sample null rejection rates.

Table 4

Estimated null rejection probabilities, with $N = 40$ and nominal significance level of 5%

	$\sigma^2 = 1$		$\sigma^2 = 10$		$\sigma^2 = 100$	
	RES	FPV	RES	FPV	RES	FPV
Case 1A	5.2	5.6	5.0	4.6	4.6	4.0
Case 1B	4.8	4.9	4.9	4.4	4.7	4.1
Case 1C	5.2	5.3	5.0	4.5	4.6	4.0
Case 2A	5.2	5.0	5.3	4.7	5.1	4.4
Case 2B	5.0	4.6	5.4	4.7	5.2	4.6
Case 2C	4.8	4.4	5.0	4.4	5.1	4.5
Case 3A	5.2	5.6	5.2	4.9	5.1	4.5
Case 3B	4.9	5.1	4.9	4.5	4.9	4.4
Case 3C	5.0	5.2	4.7	4.4	4.8	4.2
Case 4A	5.4	5.9	5.2	5.2	5.0	4.6
Case 4B	4.9	5.3	4.9	4.5	4.8	4.2
Case 4C	4.8	5.2	5.1	4.9	5.1	4.5
Case 5A	5.2	5.1	5.2	4.6	5.0	4.3
Case 5B	5.1	4.8	5.3	4.7	5.0	4.3
Case 5C	4.8	4.6	5.0	4.4	5.1	4.4
Case 6A	5.2	5.5	5.1	4.7	5.0	4.3
Case 6B	5.1	5.1	4.8	4.4	4.7	4.2
Case 6C	5.0	5.0	5.0	4.6	4.9	4.2

Note: The case codes given in the first column are derived from the codes of Table 1.

Findings concerning estimates obtained under the null hypothesis have implications for comparisons of estimates derived under alternative hypotheses. In order to make sensible comparisons of power, there should not be important differences in estimates of null rejection probabilities. There is strong evidence that, when the null hypothesis is true, MUR rejects less frequently than RES, RUR and FPV, all three of which have estimates that are closer to desired levels than those for MUR; see Table 3. It is therefore not surprising that MUR fails to detect serial correlation as frequently as the other tests when nonzero coefficients are used in (17). Given the arguments of Horowitz and Savin (2000), it was decided to exclude MUR from power comparisons, rather than to attempt to "size-correct" this test. Estimates for the remaining tests, viz., RES, RUR and FPV, are reported in Tables 5-7. These estimates are representative of the full set derived with various serial correlation models that are special cases of (17).

Table 5 contains results for regression models with errors generated by

$$(1 - 0.7B + 0.17B^2 - 0.017B^3 + 0.0006B^4)u_t = \epsilon_t, \epsilon_t iid(0,1), \tag{19}$$

which has the same AR(4) structure as the version of (8) used to generate residuals for implementing the unrestricted bootstrap test RUR. The polynomial in B used in (19) can be factorized as

$$(1 - 0.3B)(1 - 0.2B)(1 - 0.1B)^2.$$

Tables 6 and 7 contain estimates for regression models with errors generated by schemes that are not special cases of the AR(4) model; so that the version of (8) used to obtain RUR is misspecified.

Estimates in Table 6 are calculated from simulated data obtained with the serial correlation model

$$(1-0.3B)(1-0.3B^4)u_t = (1-0.3B-0.3B^4+0.09B^5)u_t = \epsilon_t, \epsilon_t iid(0,1), \tag{20}$$

which is a restricted version of an AR(5) model. For the cases of Table 7, errors are generated by

$$u_t = (1+0.3B^4)\epsilon_t, \epsilon_t iid(0,1), \tag{21}$$

i.e., by a simple MA(4) scheme. The coefficients of these and the other serial correlation models are selected by trial and error with the aim of avoiding power estimates that are close either to nominal significance levels or to 100% in all cases.

Table 5

Estimated rejection probabilities for the error model
$(1-0.7B+0.17B^2-0.017B^3+0.0006B^4)u_t = \epsilon_t, \epsilon_t iid(0,1)$,
with $N = 80$ and nominal significance level of 10%

	RES	RUR	FPV
Case 1A	55.6	55.2	55.8
Case 1B	58.2	58.0	57.7
Case 1C	56.8	56.6	56.6
Case 2A	69.2	69.2	68.8
Case 2B	70.3	70.3	69.4
Case 2C	69.7	69.8	69.2
Case 3A	73.7	73.7	74.2
Case 3B	73.8	73.9	73.7
Case 3C	73.8	73.9	73.8
Case 4A	93.8	94.0	94.2
Case 4B	94.3	94.4	94.4
Case 4C	93.8	93.8	93.9
Case 5A	79.0	79.0	78.9
Case 5B	79.5	79.6	78.9
Case 5C	79.5	79.6	79.2
Case 6A	52.9	52.4	52.9
Case 6B	53.9	53.7	53.5
Case 6C	53.2	53.1	53.1

Notes: a. The case codes given in the first column are derived from the codes of Table 1.
b. The estimates of this Table are obtained with 24956 $\geqslant$ # replications $\geqslant$ 23433.

It is clear from Table 5 that, whatever the combination of (α_1, α_2) and the error distribution, differences between power estimates are small and do not reveal a consistent ranking of RES, RUR and FPV. Consequently the results do not suggest that the unrestricted (fixed alternative hypothesis) bootstrap test RUR has better power than the restricted (null hypothesis) bootstrap test RES. The findings derived from Table 5 are corroborated by the estimates of Table 6 and Table 7.

Table 6

Estimated rejection probabilities for the error model

$(1 - 0.3B - 0.3B^4 + 0.09B^5)u_t = \epsilon_t, \epsilon_t iid(0, 1)$,

with $N = 80$ and nominal significance level of 10%

	RES	RUR	FPV
Case 1A	59.3	59.3	59.8
Case 1B	59.8	60.0	59.7
Case 1C	58.7	58.7	58.6
Case 2A	67.0	67.1	66.7
Case 2B	68.1	68.2	67.3
Case 2C	67.3	67.5	66.7
Case 3A	62.7	62.7	63.3
Case 3B	63.8	63.7	63.6
Case 3C	62.7	62.9	62.9
Case 4A	66.8	67.0	67.7
Case 4B	67.4	67.4	67.6
Case 4C	67.1	67.3	67.6
Case 5A	68.1	68.2	68.0
Case 5B	69.2	69.3	68.5
Case 5C	68.6	68.7	68.1
Case 6A	58.4	58.2	58.6
Case 6B	59.0	59.0	58.9
Case 6C	58.3	58.3	58.2

Notes: a. The case codes given in the first column are derived from the codes of Table 1.

b. The estimates of this Table are obtained with 24947 $\geqslant$ # replications $\geqslant$ 24240.

Table 7

Estimated rejection probabilities for the error model

$u_t = (1 - 0.3B^4)\epsilon_t, \epsilon_t iid(0, 1),$

with $N = 80$ and nominal significance level of 10%

	RES	RUR	FPV
Case 1A	46.2	46.2	46.7
Case 1B	46.2	46.2	45.9
Case 1C	46.1	46.1	46.0
Case 2A	47.5	47.6	47.0
Case 2B	47.7	47.8	46.8
Case 2C	47.9	47.9	47.0
Case 3A	47.1	47.2	47.5
Case 3B	47.2	47.2	46.9
Case 3C	47.5	47.3	47.2
Case 4A	45.8	45.8	46.5
Case 4B	46.4	46.3	46.2
Case 4C	45.7	45.5	45.8
Case 5A	46.4	46.5	46.2
Case 5B	46.8	47.0	45.9
Case 5C	47.0	47.0	46.3
Case 6A	47.0	47.2	47.5
Case 6B	47.2	47.3	47.0
Case 6C	47.2	47.1	47.2

Notes: a. The case codes given in the first column are derived from the codes of Table 1.

b. The estimates of this Table are obtained with 24857 $\geqslant$ # replications $\geqslant$ 23812.

5 Conclusions

Evidence concerning the behaviour of tests for serial correlation has been obtained using Monte Carlo designs that are more general in terms of dynamic specification, error distribution and alternative hypothesis than those previously employed. The F-test approach of Kiviet (1986) provides an appropriate asymptotic test in an intuitively appealing and simple fashion. However, the evidence reported above indicates that the use of an F-distribution as a source of approximate critical values can lead to an undersized test.

Three approaches to bootstrapping the F-statistic form of the Lagrange multiplier statistic in order to improve control of finite sample significance levels have been examined. The approaches differ in the treatment of the parameters of the serial correlation model, which can be treated as fixed constants, as drifting towards zero values as the sample size increases, or set equal to zero values.

The unrestricted bootstrap test RUR of Rayner (1993), based upon a fixed alternative, is vulnerable to estimation problems that affect its availability. Instrumental variable estimators of the regression model are needed for consistency under fixed alternatives and sample values of these estimators can quite frequently be at odds with the requirement that the bootstrap data generation process be dynamically stable. It is sometimes argued that unrestricted bootstrap tests can enjoy superior power but the results obtained in the Monte Carlo experiments with serially correlated errors do not support this view.

The bootstrap test MUR of Mantalos (2003) is derived by considering local, rather than fixed, alternatives. By using the local equivalence of the test model and the corresponding serial correlation alternative, it is possible to justify the use of OLS, as opposed to instrumental variables, in the unrestricted bootstrap with serially correlated errors. However, the results obtained show that asymptotic local equivalence can provide a poor approximation in finite samples and MUR can be seriously undersized. This failing under the null hypothesis leads to relatively low rejection rates in the presence of serially correlated errors.

A restricted (null hypothesis) bootstrap is used for the test RES. This test is available with very high frequency in the experiments and gives very good control of finite sample significance levels. Moreover, the power estimates for RES match those of other tests. The overall conclusion that is drawn from the simulation evidence is, therefore, that a simple restricted bootstrap should be used when testing for serial correlation after OLS estimation of models with lagged dependent variables as regressors.

The emphasis in this study has been on how best to use a single level of bootstrapping to obtain reliable and effective tests for serial correlation in dynamic regression models. The results in Beran (1988) suggest that further improvements might be derived by using a double bootstrap approach. However, the double bootstrap described in Beran (1988) has a relatively high computational cost. In recent work, Davidson and MacKinnon propose a fast double bootstrap (FDB) that is much cheaper to implement; see Davidson and MacKinnon (2006). Monte Carlo results on using a FDB test for serial correlation in a dynamic regression model are reported in Davidson and MacKinnon (2006). These Monte Carlo results are obtained from experiments in which only the one-period lag of the dependent variable is combined with strictly exogenous variables to form the regressor set and the alternative error model is a first-order autoregression. It would be interesting to examine the application of FDB methods in the context of the more general models described in Section 3 above.

Acknowledgement I am grateful to A.R. Tremayne for his constructive and helpful comments.

References

Beran, R., 1988. Prepivoting test statistics: a bootstrap view of asymptotic refinements. Journal of the American Statistical Association 83, 687-697.

Breusch, T.S., 1978. Testing for autocorrelation in dynamic linear models. Australian Economic Papers 17, 334-355.

Breusch, T.S., Godfrey, L.G., 1981. A review of recent work on testing for autocorrelation in dynamic simultaneous models. In Currie, D., Nobay, R., Peel, D. (eds.). Macroeconomic analysis: essays in macroeconomics and econometrics. Croom Helm, London.

Davidson, R., MacKinnon, J.G., 1981. Several tests for model specification in the presence of alternative hypotheses. Econometrica 49, 781-793.

Davidson, R., MacKinnon, J.G., 2004. Econometric Theory and Methods. Oxford University Press, Oxford.

Davidson, R., MacKinnon, J.G., 2006. Improving the reliability of bootstrap tests with the fast double bootstrap. Computational Statistics and Data Analysis, to appear.

Dezhbakhsh, H, 1990. The inappropriate use of serial correlation tests in dynamic linear models. Review of Economics and Statistics 72, 126-132.

Dezhbakhsh, H., Thursby, J.G., 1995. A Monte Carlo comparison of tests based upon the Durbin-Watson statistic with other autocorrelation tests in dynamic models. Econometric Reviews 14, 347-366.

Durbin, J., 1970. Testing for serial correlation in least squares regression when some of the regressors are lagged dependent variables. Econometrica 38, 410-421.

Durbin, J., Watson, G.S., 1950. Testing for serial correlation in least squares regression I. Biometrika 37, 409-428.

Durbin, J., Watson, G.S., 1951. Testing for serial correlation in least squares regression II. Biometrika 38, 159-178.

Eastwood, A., Godfrey, L.G., 1992. The properties and constructive use of misspecification tests for multiple regression models. In: Godfrey, L.G. (ed.). The implementation and constructive use of misspecification tests in econometrics. Manchester University Press, Manchester, 109-175.

van Giersbergen, N.P.A., Kiviet, J.F., 2002. How to implement the bootstrap in static or stable dynamic regression models: test statistics versus confidence region approach. Journal of Econometrics 108, 133-156.

Godfrey, L.G., 1978. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. Econometrica 46, 1293-1301.

Godfrey, L.G., 1981. On the invariance of the Lagrange multiplier test with respect to certain changes in the alternative hypothesis. Econometrica 49, 1443-1455.

Godfrey, L.G., 1988. Misspecification tests in econometrics. Cambridge University Press, Cambridge.

Godfrey, L.G., 1998. Tests for non-nested regression models: some results on small sample behaviour and the bootstrap. Journal of Econometrics 84, 59-74.

Godfrey, L.G., Tremayne, A.R., 1988. Misspecification tests for univariate time series models and their applications in econometrics. Econometric Reviews 7, 1-42.

Godfrey, L.G., Tremayne, A.R., 2005. The wild bootstrap and heteroskedasticity-robust tests for serial correlation in dynamic regression models. Computational Statistics and Data Analysis 49, 377-395.

Horowitz, J.L., 1994. Bootstrap-based critical values for the information-matrix test. Journal of Econometrics 61, 395-411.

Horowitz, J.L., Savin, N.E., 2000. Empirically relevant critical values for hypothesis tests: a bootstrap approach. Journal of Econometrics 95, 375-389.

Jeong, J., Chung, S., 2001. Bootstrap tests for autocorrelation. Computational Statistics and Data Analysis 38, 49-69.

Kiviet, J.F., 1986. On the rigour of some specification tests for modelling dynamic relationships. Review of Economic Studies 53, 241-262.

Lee, S.M.S., Young, G.A., 2005. Parametric bootstrapping with nuisance parameters. Statistics and Probability Letters 71, 143-153.

MacKinnon, J.G., 2002. Bootstrap inference in econometrics. Canadian Journal of Economics 35, 615-645.

Maddala, G.S., 2001. An Introduction to Econometrics, third edition. Wiley, New York.

Mantalos, P., 2003. Bootstrapping the Breusch-Godfrey autocorrelation test for a single equation dynamic model: Bootstrapping the restricted versus unrestricted model. Monte Carlo Methods and Applications 9, 257-269.

Mantalos, P., 2005. Two sides of the same coin: Bootstrapping the restricted versus unrestricted model. Journal of Modern Applied Statistical Methods 4, 35-42.

Ramsey, J.B., 1983. Diagnostic tests as residuals analysis: perspective and comment. Econometric Reviews 2, 241-248.

Rayner, R.K., 1993. Testing for serial correlation in regression models with lagged dependent variables. Review of Economics and Statistics 75, 716-721.

Schwert, G.W., 1987. Effects of model specification on tests for unit roots in macroeconomic data. Journal of Monetary Economics 20, 73-103.

Serlin, R.C., 2000. Testing for robustness in Monte Carlo studies. Psychological Methods 5, 230-240.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1-25.

Wooldridge, J.M., 1999. Asymptotic properties of some specification tests in linear models with integrated processes. In: Engle, R.F., White, H. (eds.). Cointegration, causality and forecasting - festschrift in honour of Clive W.J. Granger. Oxford University Press, Oxford, 366-384.