THE UNIVERSITY *of York*

Discussion Papers in Economics

No. 10/22

Robust Nonnested Testing for Ordinary Least Squares Regression When Some of the Regressors are Lagged Dependent Variables

By

Leslie G Godfrey, University of York

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

11 October 2010

Robust Nonnested Testing for Ordinary Least Squares Regression When Some of the Regressors are Lagged Dependent Variables*

Abstract

The problem of testing nonnested regression models that include lagged values of the dependent variable as regressors is discussed. It is argued that it is essential to test for error autocorrelation if ordinary least squares and the associated J and F tests are to be used. A heteroskedasticity-robust joint test against a combination of the artificial alternatives used for autocorrelation and nonnested hypothesis tests is proposed. Monte Carlo results indicate that implementing this joint test using a wild bootstrap method leads to a well-behaved procedure and gives better control of finite sample significance levels than asymptotic critical values.

JEL classification codes: C12, C15, C52
Key words: nonnested models, heteroskedasticity-robust, wild bootstrap

* I am grateful to Peter Burridge and two referees for their helpful comments.

1 INTRODUCTION

The problem of testing a model in the presence of a nonnested alternative has proved to be of importance in both applied and theoretical econometric analyses; see McAleer (1995) and Pesaran and Weeks (2001) for surveys and comments. McAleer reports that, of the various methods that have been proposed for testing nonnested regression models after ordinary least squares (OLS) estimation, the J test of Davidson and MacKinnon (1981) is the one most often used by applied workers. In order to establish the asymptotic validity of the J test, Davidson and MacKinnon make the classical assumptions that all regressors are exogenous and the errors are normally and independently distributed (NID) with common variance and zero means. However, as shown in MacKinnon et al. (1983), the J test remains asymptotically valid when the errors are independently and identically distributed (IID), but not necessarily normal, and some of the regressors are lagged values of the dependent variable, provided there is dynamic stability.

Although the assumptions used in MacKinnon et al. (1983) are much weaker than those in Davidson and MacKinnon (1981), the requirement that the errors of the model under test be IID is clearly inconsistent with modern views about best practice techniques for applied work; see, e.g., Hansen (1999) and Stock and Watson (2006) in which empirical workers are urged to adopt heteroskedasticity-robust methods. Choi and Kiefer propose "robust tests that generalize the J test ... for nonnested dynamic models with unknown serial correlation and conditional heteroscedasticity"; see Choi and Kiefer (2008). Choi and Kiefer seek to obtain a robust OLS-based J test by using heteroskedasticity and autocorrelation consistent

(HAC) methods, based upon the non-standard (fixed-b) asymptotic theory for HAC tests discussed in Kiefer et al. (2000) and Kiefer and Vogelsang (2002a, 2002b, 2005). In addition to examining the J test, Choi and Kiefer obtain a HAC variant of the comprehensive model (encompassing) F test that is sometimes used with nonnested linear regressions.

It is important to note that the asymptotic theory for robust OLS-based tests in Choi and Kiefer (2008) does not apply to all types of dynamic regression models. Regression models can be referred to as being dynamic when the regressors include lagged exogenous variables and/or lagged values of the dependent variable. The strategy for testing advocated by Choi and Kiefer is appropriate when all regressors can be taken to be current or lagged values of strictly exogenous variables. However, it cannot be employed to obtain valid HAC tests of the significance of OLS estimates in the combined presence of lagged dependent variables and serially correlated errors. The root of the problem for the HAC method and the stimulus for proposing a different procedure is that standard and fixed-b asymptotics both require that the OLS estimator be consistent; for discussions of the former and latter theories, see Greene (2008) and Kiefer and Vogelsang (2005), respectively.

In general, OLS estimators will be inconsistent when the errors are autocorrelated and there are lagged values of the dependent variable in the regressors. Consequently, if OLS-based tests of nonnested regression models with lagged dependent variables are required, it is not possible to allow the presence of unspecified forms of autocorrelation and so the assumption of serial independence is vital for the asymptotic validity of such tests. Thus, the advice given in Choi and Kiefer (2008, p. 11) that "an empirical researcher need not test the existence of serial correlation" is inappropriate in such situations and such a test is instead essential when some of the regressors are lagged values of the dependent variable.

Since the consistency of OLS estimators when the regressors include lagged dependent variables requires that the model under test has the correct regression function and that its errors have no autocorrelation, the data consistency of both of these assumptions should be checked. Given that two assumptions are under test, an applied worker can use either a joint test or two separate tests. It is argued below that, in the context of the problem examined in this paper, a joint test is more appropriate. Although robustness to error autocorrelation cannot be achieved when OLS estimation is used and the regressors include lagged dependent variables, asymptotic robustness of tests to heteroskedasticity is still feasible and desirable. The joint test is, therefore, constructed using a covariance matrix that is consistent under either unspecified forms of heteroskedasticity or homoskedasticity, provided regularity conditions are satisfied. The heteroskedasticity-robust joint tests can be implemented using either asymptotic critical values or a wild bootstrap approach and Monte Carlo evidence is provided that supports the use of the latter.

The absence of error autocorrelation is a key assumption for the consistency of OLS estimators in regression models of the type examined in this paper. If, however, OLS were replaced by an appropriate instrumental variable (IV) technique, autocorrelation would not imply inconsistency of the estimators of regression coefficients. It might be argued that it would be useful to test for autocorrelation and, if the autocorrelation test were to have a statistically significant outcome, to adopt a modified version of the robust test in Choi and

Kiefer (2008) derived using an IV estimator that is consistent under unspecified forms of autocorrelation. However, statistically significant outcomes of an autocorrelation test should be interpreted as strong evidence against the null hypothesis but not as strong evidence either for the specific autocorrelation model used as the alternative in the test, or for any form of genuine error autocorrelation. As has been emphasized many times in textbooks and research articles, tests of the null of no autocorrelation can be sensitive to many types of misspecification, e.g., omitted variables. Inferences derived from estimated "autocorrelation-corrected" covariance matrices may be very misleading when the regression function is actually underspecified. The aim of this paper is to provide heteroskedasticity-consistent tests that are convenient and reliable checks for important misspecifications but, as discussed below, these tests should not be relied upon to identify which misspecifications are present.

2 MODELS AND ASSUMPTIONS

Consider two competing nonnested regression models written as

$$H_1 : y_t = x_t'\beta_1 + u_{1t}, \tag{1}$$

and

$$H_2 : y_t = z_t'\beta_2 + u_{2t}, \tag{2}$$

in which: x_t and z_t are k_1- and k_2-dimensional vectors of regressors, with β_1 and β_2 being the corresponding vectors of unknown coefficients; u_{1t} and u_{2t} are errors terms; and there are T observations, i.e., $t = 1, ..., T$. The regressors of x_t and z_t both contain at least one lagged value of y_t. It is assumed that the true model is dynamically stable. (If no lagged dependent variables were used as regressors, the OLS-based HAC methods in Choi and Kiefer, 2008, could be employed.) A referee has pointed out that, with lagged variables used as regressors, the assumption that estimation and testing are based upon T observations implies that additional data for initial values must be available.

Suppose that the validity of H_1 is to be tested using information about H_2. (Either model could be regarded as the null. The changes required when the roles of H_1 and H_2 are reversed are straightforward.) Also suppose that the autoregressive or moving average model of order m is used as the alternative to the assumption of independent errors. Given regularity conditions, any fixed value such that $\infty > m > 0$ will deliver an asymptotically valid test under the null hypothesis. However, the choice of m will affect power. It would seem reasonable to take into account the nature of the time series data being used, e.g., $m = 4$ might be used when the data are quarterly.

It is useful to introduce the following notation: $y = (y_1, ..., y_T)'$, $X = (x_1, ..., x_T)'$, $Z = (z_1, ..., z_T)'$, $u_1 = (u_{11}, ..., u_{1T})'$ and $u_2 = (u_{21}, ..., u_{2T})'$. The OLS estimation of (1) and (2) yields vectors of coefficient estimates, predicted values and residuals denoted by:

$$\begin{aligned} \hat{\beta}_1 &= (X'X)^{-1}X'y \text{ and } \tilde{\beta}_2 = (Z'Z)^{-1}Z'y; \\ \hat{y} &= X(X'X)^{-1}X'y = P_1 y \text{ and } \tilde{y} = Z(Z'Z)^{-1}Z'y = P_2 y; \\ \hat{u}_1 &= (I_T - P_1)y = M_1 y \text{ and } \tilde{u}_2 = (I_T - P_2)y = M_2 y, \end{aligned}$$

respectively. It is usually the case that H_1 and H_2 have at least one regressor in common, e.g., an intercept term. In such cases, assume, without loss of generality, that the regressor matrix of (2) can be partitioned as $Z = (G, XA)$, with $G'M_1G$ ($k_3 \times k_3$) being positive definite and A being a known $k_1 \times (k_2 - k_3)$ matrix. Thus the k_3 variables in G are specific to H_2.

The tests of this paper can be implemented using artificial alternative models. The quasi-error terms on all artificial regressions below are denoted by u_t. Two joint tests are considered. These procedures differ in the way in which information about H_2 is incorporated.

If, as in the original J test, the predicted value from OLS estimation of H_2, is used, the artificial alternative for the joint test is

$$H_{JAC} : y_t = x_t'\beta_1 + \theta\tilde{y}_t + \rho_1\hat{u}_{1t-1} + ... + \rho_m\hat{u}_{1t-m} + u_t, \tag{3}$$

in which: $\tilde{y}_t$ is a typical element of $\tilde{y}$; and $\hat{u}_{1t-j}$ is a lagged value of the residual from OLS estimation of H_1 when $(t - j) > 0$ and is set equal to zero when $(t - j) \leq 0$. A heteroskedasticity-robust of the $(1 + m)$ restrictions of $\theta = \rho_1 = ... = \rho_m = 0$ derived from the OLS estimation of (3) is, therefore, required. Let the OLS estimators of the coefficients in (3) be denoted by $\dot{\beta}_1$, $\dot{\theta}$ and $(\dot{\rho}_1, ..., \dot{\rho}_m)$.

If, rather than following the analysis in Davidson and MacKinnon (1981), the comprehensive model approach is preferred for utilizing information about H_2, (3) is replaced by

$$H_{FAC} : y_t = x_t'\beta_1 + g_t'\gamma + \rho_1\hat{u}_{1t-1} + ... + \rho_m\hat{u}_{1t-m} + u_t, \tag{4}$$

in which g_t' denotes a typical row of G. A heteroskedasticity-robust test of the $(k_3 + m)$ restrictions $\gamma_1 = ... = \gamma_{k_3} = \rho_1 = ... = \rho_m = 0$ is then to be derived. The OLS estimators for (4) are denoted by $\ddot{\beta}_1$, $\ddot{\gamma}$ and $(\ddot{\rho}_1, ..., \ddot{\rho}_m)$.

Assumptions must be made to ensure that, when the intersection null hypothesis is true, the following two results hold: (i) the OLS estimators of the artificial alternative regression are consistent and asymptotically normal; and (ii) a valid heteroskedasticity-consistent covariance matrix estimator (HCCME) is available for these estimators. Given (i) and (ii), it is possible to derive an asymptotically valid joint test using the Wald principle with the HCCME. A basic set of regularity conditions that permits heteroskedasticity-consistent inference when regressors include lagged dependent variables is provided in Hsieh (1983). However, modifications of these assumptions are required in the context of the artificial regressions (3) and (4).

Typical observation vectors for the regressors of models (3) and (4) are written as $\hat{r}_t' = (x_t', \tilde{y}_t, \hat{u}_{1t-1}, ..., \hat{u}_{1t-m})$ and $\hat{s}_t' = (x_t', g_t', \hat{u}_{1t-1}, ..., \hat{u}_{1t-m})$, respectively. It is useful to introduce two vectors of quasi-regressors r_t and s_t that can replace $\hat{r}_t$ and $\hat{s}_t$, respectively, without affecting the asymptotic validity of tests. Let $\bar{\beta}_2(\beta_1)$ denote the probability limit of $\tilde{\beta}_2$ when the joint null hypothesis is true. The vectors for quasi-regressors are then $r_t' = (x_t', z_t'\bar{\beta}_2(\beta_1), u_{1t-1}, ..., u_{1t-m})$ and $s_t' = (x_t', g_t', u_{1t-1}, ..., u_{1t-m})$. It is assumed that, when the joint null hypothesis is true, $(u_{1t}; r_t')$ and $(u_{1t}; s_t')$ satisfy A.1 to A.7 in Hsieh (1983). For example, the counterpart of A.1 in Hsieh (1983, p. 282) is the following assumption:

Assumption 1. When H_1 is the true model, the errors u_{1t} are assumed to satisfy $E(u_{1t}|r_t) = E(u_{1t}|s_t) = 0$, which excludes endogenous variables as regressors in H_1 and also excludes autocorrelated errors; see, e.g., Choi and Kiefer (2008, p. 11).

While Hsieh's theorems provide theoretical foundations for heteroskedasticity-robust tests, practitioners need evidence about the usefulness of the asymptotic theory as a guide to finite sample behaviour. Much of the available evidence is derived from Monte Carlo studies of the behaviour of a quasi-t test, i.e., from experiments in which the null hypothesis imposes a single linear restriction on the regression coefficients. Important examples of such studies are Cribari-Neto (2004), Long and Ervin (2000) and MacKinnon and White (1985). These Monte Carlo investigations indicate that there can be substantial differences between estimates of finite sample rejection probabilities under the null hypothesis and the desired (asymptotically achieved) significance levels. However, there is an asymptotically irrelevant adjustment that seems to have a marked beneficial effect on heteroskedasticity-robust t tests. This important adjustment is to use squared restricted residuals, rather than squared unrestricted residuals, in the HCCME.

The use of restricted residuals when forming the HCCME is found in simulation experiments to yield evidence that the corresponding quasi-t test is well behaved in finite samples and that its performance is quite robust to the choice of the HCCME from the usual set of asymptotically equivalent variants; see Davidson and MacKinnon (1985a) for details. Replacing unrestricted residuals by restricted residuals in the HCCME of White (1980) leads to the estimator

$$\hat{C}_{JAC} = \left[\sum_{t=1}^{T} \hat{r}_t \hat{r}_t'\right]^{-1} \left[\sum_{t=1}^{T} \hat{u}_{1t}^2 \hat{r}_t \hat{r}_t'\right] \left[\sum_{t=1}^{T} \hat{r}_t \hat{r}_t'\right]^{-1}, \tag{5}$$

in which $\hat{r}_t' = (x_t', \tilde{y}_t, \hat{u}_{1t-1}, ..., \hat{u}_{1t-m})$, when (3) is the artificial alternative, and to the estimator

$$\hat{C}_{FAC} = \left[\sum_{t=1}^{T} \hat{s}_t \hat{s}_t'\right]^{-1} \left[\sum_{t=1}^{T} \hat{u}_{1t}^2 \hat{s}_t \hat{s}_t'\right] \left[\sum_{t=1}^{T} \hat{s}_t \hat{s}_t'\right]^{-1}, \tag{6}$$

in which $\hat{s}'_t = (x'_t, g'_t, \hat{u}_{1t-1}, ..., \hat{u}_{1t-m})$, when (4) is the artificial alternative.

Heteroskedasticity-robust Wald test statistics can be defined using (5) and (6). When testing $\theta = \rho_1 = ... = \rho_m = 0$ in (3), the test statistic is

$$\tau_{JAC} = (\dot{\theta}, \dot{\rho}_1, ..., \dot{\rho}_m) \left[R_1 \hat{C}_{JAC} R'_1 \right]^{-1} (\dot{\theta}, \dot{\rho}_1, ..., \dot{\rho}_m)', \quad (7)$$

in which $R_1 = [0_1 : I_1]$, 0_1 is a $(1 + m) \times k_1$ matrix with every element equal to zero and I_1 is the $(1 + m) \times (1 + m)$ identity matrix. The asymptotic critical values for interpreting a sample value of τ_{JAC} should be taken from the $\chi^2(1 + m)$ distribution. Similarly, when testing $\gamma_1 = ... = \gamma_{k_3} = \rho_1 = ... = \rho_m = 0$ in (4), the Wald criterion is

$$\tau_{FAC} = (\ddot{\gamma}_1, ..., \ddot{\gamma}_{k_3}, \ddot{\rho}_1, ..., \ddot{\rho}_m) \left[R_2 \hat{C}_{FAC} R'_2 \right]^{-1} (\ddot{\gamma}_1, ..., \ddot{\gamma}_{k_3}, \ddot{\rho}_1, ..., \ddot{\rho}_m)', \quad (8)$$

in which $R_2 = [0_2 : I_2]$, 0_2 is a $(k_3 + m) \times k_1$ matrix with every element equal to zero and I_2 is the $(k_3 + m) \times (k_3 + m)$ identity matrix. Asymptotic critical values for τ_{FAC} come from the $\chi^2(k_3 + m)$ distribution.

A referee has suggested that asymptotically valid heteroskedasticity-robust tests of the joint null hypothesis of correct mean specification and no error autocorrelation could be obtained by replacing $\hat{C}_{JAC}$ in (7) and $\hat{C}_{FAC}$ in (8) by the corresponding covariance matrix estimators derived from the formulae employed in Choi and Kiefer (2008). These OLS-based formulae give the covariance matrix estimator as the sum of two matrices: the first matrix is a White-type HCCME, as used above, and the second is intended to correct for autocorrelation when all regressors are exogenous (see, e.g., Greene, 2008, p.643). The calculation of the second matrix is relatively complicated, involving the sample autocovariances of cross-products of regressors and residuals, along with the choice of a kernel function and associated bandwidth parameter. More importantly, the null hypothesis specifies that there is no autocorrelation; so that the Choi-Kiefer method includes the estimation of terms that are known to be zero under the null hypothesis. The use of the Choi-Kiefer type of covariance matrix estimator, therefore, requires the applied researcher to take on the cost of the task of choosing the bandwidth and kernel function in order to end up in the position of basing heteroskedasticity-robust tests upon a covariance matrix estimator that is asymptotically inefficient relative to the HCCME. (Note that the OLS-based covariance matrix estimators in Choi and Kiefer, 2008, are not HAC for models of the type discussed here in which lagged dependent variables are used as regressors.) In the context of this paper, there is no obvious reason why the applied worker testing the joint null hypothesis would want to undertake the calculation of the "autocorrelation-correction" matrix and the use of the White-type covariance matrix estimators in (5) and (6) is preferred.

3 TEST PROCEDURES

The discussion in this section is in two parts. First, the choice between a joint test and separate tests is examined. Second, a bootstrap method for implementing heteroskedasticity-robust tests is described.

3.1 JOINT OR SEPARATE TESTS?

Whenever checking for misspecification involves testing several zero restrictions on the coefficients of an artificial alternative regression model, there is the choice between using a single joint test or a collection of separate tests. Under the null hypothesis, the joint test has the advantage that, in general, its significance level can be controlled (at least asymptotically) in a straightforward way. In contrast, given the unknown dependencies between the separate test statistics, the overall significance level associated with the collection of separate tests cannot be controlled, although an upper bound can be obtained using the Bonferroni method.

It is, of course, important to also consider the relative merits of joint and separate testing when there is misspecification. If the joint (intersection) null hypothesis is untrue, it is to be hoped that the joint test will produce a statistically significant outcome with reasonably high probability. However, such statistically significant outcomes of a joint test cannot provide information about the ways in which the null model is misspecified. If it were the case that separate tests could identify the source of misspecification, there would be an argument against the use of a joint test. However, there are good reasons in general settings and in the particular context of this paper for believing that, when the intersection null hypothesis is untrue, separate tests cannot be assumed to be reliable guides to re-specification and/or choice of an alternative estimator.

The general problem that impedes the constructive use of separate tests is that they can be sensitive to misspecifications that are not in the class of alternative models for which they were designed. For example, it is stressed in textbooks, e.g., Greene (2008, Ch. 19), that statistically significant autocorrelations of OLS residuals can be caused by misspecification of the regression model and Mizon argues that "if the null hypothesis of no serial correlation is rejected, there is not a unique alternative model to adopt, since all the test result has established is that the present model is inadequate, probably by having an inappropriate dynamic specification"; see Mizon (1995). After an analysis of the asymptotic behaviour of separate test statistics, it is concluded in Davidson and MacKinnon (1985b) that a separate test statistic "can indeed tell us that a model is wrong, but by itself it can never tell us why."

In terms of the specific framework of this paper, the separate autocorrelation test would be likely to sensitive if H_2 with independent errors were the true data generation process (DGP) and H_2 contained more lagged values of the dependent variable than H_1; see the discussion of the sensitivity of autocorrelation tests to the omission of relevant lagged dependent variables provided in Davidson and MacKinnon (1985b, pp. 44-46). Turning to separate nonnested hypothesis tests, a referee has pointed out that, when the DGP is H_1 with autocorrelated errors, the inconsistency of OLS is likely to lead to non-zero probability limits of the OLS estimators of coefficients of the variables used to incorporate information about H_2 and hence to separate tests of H_1 against H_2 having rejection probabilities in excess of their significance levels.

The combination of H_1 and autocorrelated errors is outside the null hypothesis of this paper because it is being assumed that the applied worker wishes to use OLS to estimate equations with lagged dependent variables as regressors. However, this type of DGP may be of interest in an empirical study. A referee has pointed out that if an IV estimator, derived using only exogenous variables as instruments, were to be used in place of OLS, it might be possible to extend the asymptotic analysis in Choi and Kiefer (2008) to obtain IV-based HAC tests of H_1 based upon information about H_2. Discussions of the important issues of the finite sample behaviour of such IV-based procedures and the choice of instruments, kernel and bandwidth for the construction of the test statistics are beyond the scope of this paper but are interesting areas for future research.

It does, however, seem reasonable to conjecture that bootstrap methods, rather than asymptotic critical values, will be needed to get good control of finite sample significance levels of IV-based tests. The consistency of an IV estimator does not imply that bootstrap applications in finite samples will be without problems. In a study of homoskedasticity-only bootstrap tests for serial correlation for models with lagged dependent variables included as regressors, Godfrey finds that the use of IV methods that are consistent under the alternative hypothesis leads to problems in finite samples. His Monte Carlo results show that the IV estimates of regression coefficients, which serve as the parameters of the bootstrap data generation process, often fail to satisfy the conditions for dynamic stability; see Godfrey (2007).

3.2 BOOTSTRAP IMPLEMENTATION OF THE JOINT TEST

There is evidence that even the use of restricted residuals in the HCCME is not sufficient to obtain good approximations from the asymptotic theory when the null hypothesis imposes several restrictions, rather than just one. More precisely, it has been found that, when several restrictions are under test, the use of asymptotic critical values with restricted residual HCCME-based test statistics produces estimates of null hypothesis rejection probabilities that are too small; see Godfrey and Orme (2004). This point is important here, given that the J and F approaches, combined with a check for mth order autocorrelation, lead to joint tests of $(1 + m) > 1$ and $(k_3 + m) > 1$ restrictions, respectively. The failure of asymptotic critical values to provide reliable inferences in general situations of empirical interest has led several researchers to examine the use of bootstrap methods.

Results that are relevant to the implementation of bootstrap tests in conditionally heteroskedastic dynamic regression models are provided in Gonçalves and Kilian (2004). The asymptotic validity of three methods is established by Gonçalves and Kilian who also report Monte Carlo evidence on finite sample behaviour. In two of the methods, referred to as the "pairwise bootstrap" and the "fixed-design wild bootstrap", lagged values of the dependent variable that are included as regressors are treated as if they were exogenous. The third method, called the "recursive-design wild bootstrap", mimics the dynamic nature of the assumed data process and allows for conditional heteroskedasticity by combining the recursive bootstrap for autoregressions of Bose (1988) with the heteroskedasticity-valid wild bootstrap of, e.g., Liu (1988). Thus, if the regressors of (1) are ordered so that $x_t' = (y_{t-1}, ..., y_{t-l}; q_t')$, with q_t containing the exogenous regressors, the recursive-design wild bootstrap generates artificial data according to

$$y_t^* = (y_{t-1}^*, ..., y_{t-l}^*; q_t')\hat{\beta}_1 + u_{1t}^*, \text{ with } u_{1t}^* = \hat{u}_{1t}\epsilon_t, t = 1, ..., T, \tag{9}$$

in which required starting values are set equal to those for the actual data and the ϵ_t are IID drawings from a "pick distribution" that has zero mean, variance equal to unity and finite fourth moment.

After considering the results from their Monte Carlo experiments, Gonçalves and Kilian conclude that, of the three methods that they examine, the recursive-design wild bootstrap of (9) "seems best suited for applications in empirical macroeconomics"; see Gonçalves and Kilian (2004, p. 106). In general, when discussing Monte Carlo evidence, Gonçalves and Kilian use the results obtained with the ϵ_t of (9) being distributed as standard normal. They remark, however, that these results are robust to changes of the pick distribution in which the standard normal is replaced by either of two well-known two-point distributions. One of these two-point distributions is proposed in Mammen (1993) and the other is the Rademacher distribution defined by

$$\Pr(\epsilon_t = -1) = \Pr(\epsilon_t = 1) = 0.5. \tag{10}$$

Several researchers have reported evidence that supports the use of (10); see Davidson et al. (2007), Davidson and Flachaire (2008) and Flachaire (2005) for results for models with exogenous regressors and Godfrey and Tremayne (2005) for findings about models in which lagged dependent variables are included in the regressors. The tests of Section 2 are, therefore, implemented using (9) with (10). The algorithm for carrying out the required heteroskedasticity-robust joint test can then be described as follows.

Step 1. Estimate (1) and (2) by using OLS with the actual data $\Delta = (y; X; Z)$ to obtain residuals from the former model and predicted values from the latter model if the J test approach is to be used.

Step 2. According to the choice of general approach to testing H_1 in the presence of information about H_2, while jointly testing for autocorrelation, estimate either (3) or (4) by OLS and obtain the corresponding HCCME as defined in either (5) or (6). Use this HCCME to calculate the heteroskedasticity-robust Wald statistic from the actual data Δ. Let the sample value of the statistic be denoted by τ.

Steps 3 to 5 are repeated B times. A bootstrap sample is produced and used to calculate the bootstrap counterpart of τ each time. In the descriptions of Steps 3 to 5, j is used to denote a typical repetition, so $j = 1, ..., B$.

Step 3. Generate a sequence of artificial observations $y^*_{(j)t}$, $t = 1, ..., T$, using the recursive wild bootstrap scheme that consists of (9) and (10). Values of exogenous regressors in (1) and (2) are held fixed over bootstrap samples and the terms $\{y^*_{(j)t}\}$, combined with any required actual starting values, provide the specified lagged dependent variables for regressor sets; so that the bootstrap sample $\Delta^*_{(j)} = (y^*_{(j)}; X^*_{(j)}; Z^*_{(j)})$ is now available.

Step 4. Perform the calculations of Step 1 with actual data Δ replaced by bootstrap data $\Delta^*_{(j)}$.

Step 5. Perform the calculations of Step 2 with actual data Δ replaced by bootstrap data $\Delta^*_{(j)}$. Let the bootstrap test statistic be denoted by $\tau^*_{(j)}$.

After Steps 3 to 5 have been carried out B times, the bootstrap p-value of the observed test statistic τ can be calculated as the proportion of the bootstrap test statistics that are at least as large as the actual value, i.e., $\hat{p}^*(\tau) = \#(\tau^*_{(j)} \geq \tau)/B$. The intersection null hypothesis of interest is rejected if $\hat{p}^*(\tau) \leq \alpha$, in which α denotes the desired significance level.

4 MONTE CARLO DESIGN

Following, e.g., Delgado and Stengos (1994) and Fan and Li (1995), the Monte Carlo experiments are based upon designs in Godfrey and Pesaran (1983). Each of the two models has y_{t-1} as the first regressor and all other regressors are exogenous. As in the previous sections, it is assumed that the first model is under test.

4.1 Experiments in which the joint null hypothesis is true

When calculating rejection frequencies that correspond to finite sample significance levels of heteroskedasticity-robust joint tests, data are generated under H_1 according to

$$y_t = \psi_1 x_{1t} + \sum_{i=2}^{k_1} x_{it} + u_{1t}, \text{ with } x_{1t} = y_{t-1}, t = 1, .., T, \tag{11}$$

$\psi_1 = (0.3, 0.5, 0.7)$, and the competing model H_2 is written as

$$y_t = \sum_{i=1}^{k_2} \pi_i z_{it} + u_{2t}, \text{ with } z_{1t} = y_{t-1}, t = 1, .., T. \tag{12}$$

The numbers of regressors in the experiments are $(k_1, k_2) = \{(3,3), (3,5), (5,3), (5,5)\}$. Note that k_1 and k_2 in (11) and (12) correspond to $k_0 + 1$ and $k_1 + 1$, respectively, in the notation of Godfrey and Pesaran (1983). The terms $x_{it}, i = 2, .., k_1$, in (11) are exogenous and are obtained as standard normal variates that are independent over i and t. The last $k_2 - 1$ regressor values in (12) are given by $z_{it} = \lambda x_{it} + \xi_{it}$, $i = 2, ..., \min(k_1, k_2)$ and, if $k_1 < k_2$, $z_{it} = \xi_{it}$, $i = k_1 + 1, ..., k_2$, with the ξ_{it} being standard normal variates that are independent over i and t. The value of λ is determined by fixing the positive population correlation between x_{it} and z_{it}, which is denoted by ρ, with $\rho^2 = (0.3, 0.9)$.

The errors u_{1t} are independently and normally distributed with zero means. (Using transformations of drawings from either $\chi^2(5)$ or $t(5)$ distributions to obtain the errors did not alter the results in important ways.) Conditional variances are obtained in three ways. First, there is conditional homoskedasticity with $\sigma_{1t}^2 = Var(u_{1t}|x_t, z_t) = \sigma_1^2$, for all t, with σ_1^2 determined via $\sigma_1^2 = ((k_1 - 1)(1 - R_1^2))/(R_1^2 - \psi_1^2)$, in which $R_1^2 = (0.5, 0.8)$ is the population coefficient of determination for H_1. Second, the GARCH model is used with

$$\sigma_{1t}^2 = c_0 + 0.1u_{1t-1}^2 + 0.8\sigma_{1t-1}^2, \tag{13}$$

in which c_0 is selected to yield an unconditional variance equal to σ_1^2 as defined for the case of homoskedastic errors. The coefficients of u_{1t-1}^2 and σ_{1t-1}^2 in (13) are similar to those reported in empirical work. Third, a model based upon the HET1 scheme in Andrews (1991) is used, with $\sigma_{1t}^2 = \sigma_1^2 x_{2t}^2$, in which σ_1^2 is calculated as in the other two schemes.

There are $T = (40, 80)$ observations, so it seems reasonable to think in terms of quarterly data and to use $m = 4$ in the artificial regressions that correspond to (3) and (4). The joint test that uses the J test variable, therefore, tests 5 restrictions. There are $(k_2 - 1)$ variables that are specific to (12) and so the F-based joint test using the appropriate special case of (4) tests $(k_2 + 3)$ zero restrictions for $k_2 = 3, 5$. The desired significance level is $\alpha = 5$ per cent. The corresponding finite sample estimates are obtained using 25,000 replications. All asymptotic critical values are taken from χ^2 distributions. The wild bootstrap versions of the joint tests are carried out with $B = 400$ bootstrap samples.

4.2 Experiments in which the joint null hypothesis is untrue

Three departures from the model of the joint null hypothesis are used: first, the DGP is H_2 with independent errors; second, the DGP is H_1 with autocorrelated errors; and third, the DGP is H_2 with autocorrelated errors. In the first set of experiments, data are generated under H_2 according to

$$y_t = \psi_2 z_{1t} + \sum_{i=2}^{k_2} z_{it} + u_{2t}, \text{ with } z_{1t} = y_{t-1}, t = 1, .., T, \tag{14}$$

$\psi_2 = (0.3, 0.5, 0.7)$, and the competing model is

$$y_t = \sum_{i=1}^{k_1} \omega_i x_{it} + u_{1t}, \text{ with } x_{1t} = y_{t-1}, t = 1, .., T. \tag{15}$$

Generation of exogenous regressor values is exactly the same as when (11) is the Monte Carlo DGP. Also the coefficients used to determine conditional means are the same. Conditional variance models only differ in non-trivial ways from those for experiments designed to provide estimates of significance levels because the expression for the unconditional error variance must be altered. For example, in the HET1-type scheme for u_{2t} of (14), $\sigma_{2t}^2 =$

$Var(u_{2t}|x_t, z_t) = \sigma_2^2 z_{2t}^2$ and the parameter σ_2^2 is determined by $\sigma_2^2 = \vartheta_z^2(1 - R_2^2)/(R_2^2 - \psi_1^2)$, in which: $\vartheta_z^2 = (k_2 - 1)(1 + \lambda^2)$ if $k_1 \geq k_2$; $\vartheta_z^2 = \{(k_2 - 1) + (k_1 - 1)\lambda^2\}$ if $k_1 < k_2$; and $R_2^2 = (0.5, 0.8)$ is the population coefficient of determination for H_2. The number of replications used to estimate power is 10,000; so that the maximum standard error of the estimator of the rejection probability is 0.5 per cent, which seems adequate for practical purposes.

The second set of experiments, i.e., H_1 with autocorrelated errors, is used to examine the impact of the choice of m on power. In order to focus on the importance of the order of the autocorrelation alternative, the nonnested alternative (12) is used with $k_2 = 2$, implying the equivalence of F and J test variables (and invariance of results with respect to ρ^2). Joint tests are obtained with $m = 1$ and $m = 4$, with the errors being generated using a generalization of a scheme in Andrews (1991). More precisely, the errors are obtained from

$$v_t = \sum_{j=1}^{4} \phi_j v_{t-j} + a_t, \quad a_t \;\; NID(0, \sigma_a^2), \tag{16}$$

and $u_{1t} = |x_{2t}|v_t$, with v_t in (16) being stationary. Two versions of (16) are employed. The first of these schemes has $\phi_1 = 0.75$ and $\phi_j = 0$, $j = 2, 3, 4$; so that using $m = 4$ implies that irrelevant test variables have been used and power losses relative to the test based on $m = 1$ are expected. The second version of (16) has $\phi_1 = 0.7$, $\phi_2 = -0.17$, $\phi_3 = 0.017$ and $\phi_4 = -0.0006$; so that (1 - 0.3$\mathcal{L}$)(1 - 0.2$\mathcal{L}$)(1 - 0.1$\mathcal{L}$)$^2 v_t = a_t$, in which $\mathcal{L}$ denotes the lag operator. There is no generally valid prediction about the ranking by power of the tests based upon $m = 1$ and $m = 4$ when this second autocorrelation scheme is part of the Monte Carlo DGP. The value of σ_a^2 in (16) is selected by trial and error to avoid rejection frequencies of the false intersection null hypothesis that are close either to the nominal significance level or to 100 per cent. Required initial values are set equal to zero and $T+51$ observations are generated. The first 50 observations are dropped to reduce the impact of the fixed initial values and the 51st observation is used for $t=0$.

The third set of experiments uses, as the basic form of the DGP, H_2 of (14), with $k_2=2$ and errors given by $u_{2t}=|z_{2t}|v_t$, in which v_t is autocorrelated according to (16). As in the experiments in which the DGP is H_2 with independent errors, results depend upon the value of ρ^2. The two sets of values of the coefficients $\{\phi_j\}$ in (16) are the same as in the second set of experiments, as are the strategy for choosing σ_a^2 and the treatment of initial values. Joint tests of the claim that the correct specification is H_1 with independent errors are conducted by using both $m=1$ and $m=4$.

5 Monte Carlo Results

To save space, results are only discussed for experiments with conditional heteroskedasticity of the HET1-type. The general findings that the recursive wild bootstrap leads to a well-behaved procedure and gives better control of finite sample significance levels than asymptotic critical values are not altered when either GARCH or homoskedastic errors are used.

5.1 Results when null hypothesis is true

Tables 1 and 2 contain a representative sample of the results obtained when the intersection null hypothesis that H_1 is valid and has non-autocorrelated errors is true. The results for the combination of the artificial regressors of the J test and the Breusch-Godfrey test are given in Table 1 and those for the combination of the F-type test variables and the Breusch-Godfrey variables are given in Table 2. For each set of design parameters, rejection frequencies are calculated using asymptotic and recursive wild bootstrap critical values. With 25,000 replications, the standard error of the proportion of rejections would be $\sqrt{5(95)/25,000} = 0.14$ per cent (approximately) if the true finite sample significance level were equal to the desired value of 5 per cent.

TABLE 1

H_1 versus H_{JAC}: rejection frequencies when null hypothesis is true, with desired level of 5 per cent, for asymptotic and recursive wild bootstrap tests

R_1^2	ρ^2	ψ_1	T	$(k_1, k_2) =$	*(3, 3)*		*(3, 5)*		*(5, 3)*		*(5, 5)*	
0.5	0.3	0.3	40		1.5	**4.5**	2.2	**4.6**	2.0	**4.4**	2.5	**4.3**
0.5	0.3	0.5	40		1.6	**4.5**	2.5	**4.6**	2.2	**4.5**	2.7	**4.5**
0.5	0.3	0.7	40		2.1	**4.8**	3.6	**4.7**	2.6	**4.7**	4.0	**4.7**
0.5	0.9	0.3	40		1.6	**4.5**	2.3	**4.5**	2.0	**4.5**	2.0	**4.4**
0.5	0.9	0.5	40		1.6	**4.6**	2.4	**4.5**	2.1	**4.5**	2.1	**4.4**
0.5	0.9	0.7	40		1.8	**4.6**	2.7	**4.6**	2.0	**4.3**	2.3	**4.5**
0.5	0.3	0.3	80		2.9	**5.2**	3.5	**5.0**	3.2	**5.1**	3.6	**4.9**
0.5	0.3	0.5	80		3.0	**5.1**	4.0	**5.0**	3.4	**5.0**	3.9	**5.0**
0.5	0.3	0.7	80		3.6	**5.2**	6.7	**5.0**	3.9	**5.0**	6.4	**5.4**
0.5	0.9	0.3	80		2.9	**5.2**	3.8	**4.8**	3.1	**5.0**	3.1	**4.9**
0.5	0.9	0.5	80		2.9	**5.2**	4.0	**4.8**	3.1	**5.0**	3.1	**4.8**
0.5	0.9	0.7	80		3.0	**5.2**	4.9	**4.8**	3.1	**4.7**	3.4	**5.0**

Notes: The first figure of each pair is derived using asymptotic critical values and the second figure (given in **bold**) is derived using the recursive wild bootstrap method. All results are reported as percentages, rounded to one decimal place.

Consider first the strategy in which use is made of the test variable of Davidson and

MacKinnon (1981) and, in the notation of Section 2, H_1 is tested against H_{JAC}. Experiments based upon the Monte Carlo DGPs described above produce strong evidence against the claim that asymptotic theory provides a good approximation to the finite sample behaviour of test statistics. When $T = 40$, the majority of rejection frequencies obtained using asymptotic critical values do not exceed half of the required rejection probability of 5 per cent. Increasing the sample size and using $T = 80$ improves the approximation provided by asymptotic theory, but there is still evidence that marked under-rejection is common. Some results for cases with $R_1^2 = 0.5$ and $\psi_1 = 0.7$ are larger than those associated with other cases. However, $\psi_1 = 0.7$ implies that $\psi_1^2 = 0.49$; so that, when $R_1^2 = 0.5$, the exogenous variables $x_{it}, i = 2, ..., k_1$, have very little impact on the population goodness of fit. If these exogenous variables were actually irrelevant, it would follow that: $R_1^2 = \psi_1^2 = 0.49$; (11) would be nested in (12); and there would be a breakdown of the standard asymptotic theory of the J test.

In contrast to the results derived from asymptotic critical values, the rejection frequencies in Table 1 that are obtained using the recursive wild bootstrap approach suggest much better control of finite sample rejection probabilities. When $T = 40$, there is clearly a small degree of under-rejection, with rejection frequencies fluctuating around 4.5 per cent, rather than the nominal value of 5 per cent. This good performance is improved by increasing the sample size to $T = 80$, with almost all of the rejection frequencies then being within 1.5 standard errors of the nominal value and the general level of agreement being very good as judged by, e.g., the criteria in Serling (2000).

The results for the heteroskedasticity-robust joint test derived by combining the artificial comprehensive model for the F-test with the artificial regression of the Breusch-Godfrey test are presented in Table 2. The derivation of H_1 as the null model when H_{FAC} is the alternative model requires more restrictions to be imposed than when H_{JAC} plays the latter role. It is, therefore, not surprising that Table 2 shows poorer approximations from asymptotic theory than are seen in Table 1. Table 2 contains clear evidence that the true significance levels implied by using asymptotic critical values are much smaller than the desired level of 5 per cent. The rejection frequencies for $T = 40$ in Table 2 are in the range 0.7 to 1.8 per cent and those for $T = 80$ are between 1.9 and 3.0 per cent. Fortunately, as observed in Table 1, the recursive wild bootstrap provides much more useful approximations.

TABLE 2

H_1 versus H_{FAC}: rejection frequencies when null hypothesis is true, with desired level of 5 per cent, for asymptotic and recursive wild bootstrap tests

R_1^2	ρ^2	ψ_1	T	$(k_1, k_2) =$	*(3, 3)*		*(3, 5)*		*(5, 3)*		*(5, 5)*	
0.5	0.3	0.3	40		1.2	**4.4**	0.7	**4.6**	1.6	**4.3**	1.0	**4.0**
0.5	0.3	0.5	40		1.2	**4.3**	0.8	**4.5**	1.8	**4.5**	1.0	**4.2**
0.5	0.3	0.7	40		1.3	**4.5**	0.7	**4.5**	1.8	**4.4**	0.9	**4.1**
0.5	0.9	0.3	40		1.2	**4.4**	0.7	**4.6**	1.6	**4.3**	1.0	**4.0**
0.5	0.9	0.5	40		1.2	**4.3**	0.8	**4.5**	1.8	**4.5**	1.0	**4.2**
0.5	0.9	0.7	40		1.3	**4.5**	0.7	**4.5**	1.8	**4.4**	0.9	**4.1**
0.5	0.3	0.3	80		2.6	**5.2**	2.0	**4.9**	3.0	**5.2**	2.4	**4.9**
0.5	0.3	0.5	80		2.6	**5.2**	2.0	**5.0**	2.9	**5.0**	2.4	**4.8**
0.5	0.3	0.7	80		2.6	**5.3**	2.0	**4.8**	2.7	**4.7**	2.3	**4.7**
0.5	0.9	0.3	80		2.6	**5.2**	2.0	**4.9**	3.0	**5.2**	2.4	**4.9**
0.5	0.9	0.5	80		2.6	**5.2**	2.0	**5.0**	2.9	**5.0**	2.4	**4.8**
0.5	0.9	0.7	80		2.6	**5.3**	2.0	**4.8**	2.7	**4.7**	2.3	**4.7**

Notes: The first figure of each pair is derived using asymptotic critical values and the second figure (given in **bold**) is derived using the recursive wild bootstrap method. All results are reported as percentages, rounded to one decimal place.

When the recursive wild bootstrap is employed in the test of H_1 against H_{FAC}, the results for $T = 40$ can reasonably be viewed as consistent with the claim that, for each case studied, the true significance level is in the range 4.4 ± 0.2 per cent. The rejection frequencies in Table 2 for cases with $T = 80$ are even closer to the desired value of 5 per cent and they are all consistent with the claim that the corresponding true significance level is in the range 5.0 ± 0.2 per cent.

5.2 Results when null hypothesis is untrue

Since the evidence obtained when the intersection null hypothesis is true suggests that recursive wild bootstrap tests of H_1 against H_{JAC} and of H_1 against H_{FAC} have similar and well-behaved significance levels in finite samples, it seems reasonable to compare their rejection frequencies when the intersection null hypothesis is false. In contrast, the corresponding tests that use asymptotic critical values are excluded because of the failure of asymptotic theory to give good control of finite sample significance levels. No attempt is made to include the asymptotic tests after using so-called size-corrections derived from Monte Carlo results because, as argued persuasively in Horowitz and Savin (2000), such corrections are not relevant to empirical research.

5.2.1 DGP is $\mathbf{H_2}$ with no autocorrelation of the errors

Consider first some results obtained by generating samples under a data process of the type H_2 with independent conditionally heteroskedastic errors, with the wild bootstrap tests being carried out with a nominal significance level of 5 per cent. The tests derived from the counterparts of (3) and (4) differ in the way in which information about the specification of H_2 is incorporated. For the models used in the Monte Carlo experiments, the test variables employed in the comprehensive model F-test method are the $(k_2 - 1)$ exogenous regressors that are specific to H_2, $k_2 = 3, 5$. The Davidson-MacKinnon approach is equivalent to weighting these regressors by the OLS estimators of the corresponding coefficients in order to obtain a single test variable, rather than $(k_2 - 1)$ test variables, $k_2 = 3, 5$. When H_2 is the true model, these OLS estimators have probability limits equal to genuine parameters of interest and the weighting is well-founded. It is not surprising that previous research has found the J-test to be more powerful than the F-test when H_1 is tested and H_2 is the true DGP; see, e.g., Godfrey (1998) in which an IID-valid residual bootstrap method is used for both tests. The results in Table 3 provide some evidence on the relative merits of these approaches to non-nested testing when they are both combined with a check for autocorrelation in a way that gives asymptotic robustness to conditional heteroskedasticity.

The results in Table 3 are for $R_2^2 = 0.8$, with other design parameters being selected to give rejection rates for the untrue intersection null hypothesis that are neither too small nor too large to provide interesting comparisons. The general features of the results are as expected. First, holding other things constant, increasing T from 40 to 80 increases rejection frequencies. Second, as the positive coefficient ψ_2 increases, other features fixed, estimates of power fall, which reflects the increasing importance of the lagged dependent variable, which is a regressor in both (14) and (15), relative to the model-specific exogenous regressors in (14). (The same general features are observed when $R_2^2 = 0.5$.)

TABLE 3

Rejection frequencies, with a nominal significance level of 5 per cent, when the intersection null hypothesis is untrue and the true model is H_2 with independent conditionally heteroskedastic errors and $R_2^2 = 0.8$

					$T = 40$		$T = 80$	
ρ^2	ψ_2	k_1	k_2	Artificial alternative is	H_{FAC}	H_{JAC}	H_{FAC}	H_{JAC}
0.9	0.3	3	3		59.7	**65.4**	94.4	**95.6**
0.9	0.5	3	3		52.1	**57.4**	89.6	**91.5**
0.9	0.7	3	3		36.3	**40.4**	72.3	**74.8**
0.9	0.3	3	5		64.6	**80.6**	98.5	**99.4**
0.9	0.5	3	5		58.4	**74.0**	96.6	**98.4**
0.9	0.7	3	5		44.0	**56.7**	86.8	**92.5**
0.9	0.3	5	3		56.1	**61.9**	93.8	**95.3**
0.9	0.5	5	3		48.3	**54.0**	88.6	**90.8**
0.9	0.7	5	3		31.9	**36.6**	69.9	**73.4**
0.9	0.3	5	5		45.2	**62.6**	91.0	**95.7**
0.9	0.5	5	5		38.6	**54.4**	85.0	**91.7**
0.9	0.7	5	5		26.3	**37.2**	65.3	**73.3**

Notes: Both tests of the untrue model H_1 are carried out using the recursive wild bootstrap method, with the Rademacher pick distribution. All results are reported as percentages, rounded to one decimal place. Results for the test that combines the J-test with the autocorrelation test are given in **bold**.

Turning to the comparison of the two joint tests τ_{JAC} and τ_{FAC}, the results in Table 3 show that using the J-type method with the lagged residuals always leads to a greater rejection frequency than the corresponding heteroskedasticity-robust joint test based on the F (comprehensive model) approach. Also, other things being equal, the shortfall of the joint test that uses the comprehensive model F approach increases when k_2 increases from 3 to 5. Neither of these outcomes is surprising, given the discussion above concerning the use and construction of test variables. However, it is important to recognize that, if H_1 and H_2 were both untrue, the F approach would yield a heteroskedasticity-robust joint test that was consistent against a wider range of alternatives than the J approach; see Mizon and Richard (1986, Section 4). The power differences observed in Table 3 can be seen as costs of some insurance against the event that both models under consideration are misspecified.

Some additional experiments are carried out to assess the effects of the coefficient of the lagged dependent variable being greater than the value used in the main set of experiments, i.e., 0.7. Setting ($\psi_1 = 0.9, R_1^2 = 0.95$) in (11) and ($\psi_2 = 0.9, R_2^2 = 0.95$) in (14) does not lead to evidence of either poor control of significance levels or low power. As anticipated from the results of Table 3, the lowest rejection frequencies for an untrue null hypothesis are observed when, in addition to the common lagged variable having a high coefficient, the exogenous regressors of the competing models are highly correlated. Table 4 contains results for $\rho^2 = 0.9$ that suggest that, under the most demanding combination of design parameters, false models can be detected with reasonable probability with $T = 40$ and rather more frequently when $T = 80$. The rejection frequencies in Table 4 also provide another illustration of the good control over the significance level that is obtained by using the wild bootstrap approach.

TABLE 4

Rejection frequencies for tests of H_1, with a nominal significance level of 5 per cent and independent conditionally heteroskedastic errors, when the lagged dependent variable has a coefficient equal to 0.9 and $\rho^2 = 0.9$

DGP is H_1 of (11)					$T = 40$		$T = 80$	
R_1^2	ψ_1	k_1	k_2	Artificial alternative is	H_{FAC}	H_{JAC}	H_{FAC}	H_{JAC}
0.95	0.9	3	3		4.6	**4.6**	5.2	**5.1**
0.95	0.9	3	5		4.3	**4.5**	4.7	**4.8**
0.95	0.9	5	3		4.3	**4.6**	5.0	**4.9**
0.95	0.9	5	5		4.1	**4.6**	4.9	**5.0**
DGP is H_2 of (14)					$T = 40$		$T = 80$	
R_2^2	ψ_2	k_1	k_2	Artificial alternative is	H_{FAC}	H_{JAC}	H_{FAC}	H_{JAC}
0.95	0.9	3	3		52.8	**57.9**	89.9	**91.7**
0.95	0.9	3	5		58.9	**74.6**	96.8	**98.6**
0.95	0.9	5	3		48.5	**54.3**	98.5	**99.4**
0.95	0.9	5	5		38.5	**54.9**	85.2	**91.8**

Notes: Both tests of the intersection null hypothesis are carried out using the recursive wild bootstrap method, with the Rademacher pick distribution. All results are reported as percentages, rounded to one decimal place. Results for the test that combines the J-test with the autocorrelation test are given in **bold**.

5.2.2 DGP is $\mathbf{H_1}$ with autocorrelation of the errors

As explained above, the nonnested alternative model H_2 with $k_2 = 2$ is used in the experiments based upon a DGP consisting of H_1 with autocorrelated errors. Thus there is only one regressor that is specific to the nonnested alternative model, implying that the F and J approaches coincide, and it is possible to focus on sensitivity to the choice of lag-length in the Breusch-Godfrey component of the artificial alternative. Table 5 contains a sample of the results concerning this sensitivity.

TABLE 5

Rejection frequencies, with a nominal significance level of 5 per cent, when the DGP is H_1 with autocorrelated and heteroskedastic errors

(a) Coefficients of (16): $\phi_1 = 0.75$; $\phi_2 = 0$; $\phi_3 = 0$; $\phi_4 = 0$; $\sigma_a^2 = 0.1$							
				$T = 40$		$T = 80$	
ψ_1	k_1	k_2	Value of m is	1	4	1	4
0.3	3	2		59.4	44.6	92.8	88.9
0.5	3	2		56.7	42.0	91.3	86.6
0.7	3	2		55.6	39.8	91.0	84.8
(b) Coefficients of (16): $\phi_1 = 0.7$; $\phi_2 = -0.17$; $\phi_3 = 0.017$; $\phi_4 = -0.0006$; $\sigma_a^2 = 0.1$							
				$T = 40$		$T = 80$	
ψ_1	k_1	k_2	Value of m is	1	4	1	4
0.3	3	2		44.5	26.7	84.9	67.4
0.5	3	2		43.4	26.0	84.9	67.0
0.7	3	2		43.1	25.7	85.4	66.8

Notes: The test of the untrue intersection null hypothesis is carried out using the recursive wild bootstrap method, with the Rademacher pick distribution. All results are reported as percentages, rounded to one decimal place.

Panel (a) of Table 5 provides an illustration of the costs of using too high a lag-length. It is clear that using $m = 4$ when $m = 1$ is appropriate can lead substantial reductions in rejection frequencies; so that there is evidence of the usual problem of irrelevant test variables causing loss of power. The results in panel (b) of Table 5 are for cases in which $m = 4$ is appropriate and so the use of $m = 1$ implies underspecification of the artificial alternative. In the evidence reported in panel (b), it is clear that underspecification has produced greater rejection rates than the correct choice. There is no generally valid result that is supported by this evidence. Using too small a value for m can, in other situations, lead to reductions relative to the correct choice. For example, the results in Godfrey and Tremayne (1988, p. 33, Table 3) illustrate that, when the true error model is a simple fourth-order autoregression, the use of an underspecified first-order alternative produces a test that is less sensitive than the test using the correct alternative and is also inferior to the test obtained from an overspecified sixth-order scheme. It is not only the number of restrictions being tested that matters but also the ability of the test variables for the assumed alternative to approximate those appropriate for the true error process. The results in panel (b) do, of course, illustrate the general dangers of assuming that errors are first-order autoregressive just because a test with $m = 1$ produces a statistically significant result.

When the DGP is H_1 with autocorrelated and heteroskedastic errors, rejection rates will depend upon the strength of autocorrelation. For example, weakening the degree of autocorrelation by using $\phi_1 = 0.4$, rather than $\phi_1 = 0.75$, in the first scheme leads to the correct choice of $m = 1$ producing rejection frequencies of about 20 per cent when $T = 40$ and about 50 per cent when $T = 80$. (The corresponding approximate figures when there is overspecification, with $m = 4$ being used, are 10 per cent when $T = 40$ and 30 per cent when $T = 80$.)

5.2.3 DGP is $\mathbf{H}_2$ with autocorrelation of the errors

In the final set of experiments, the DGP is H_2 with autocorrelated and heteroskedastic errors. When H_1 is false and H_2 has autocorrelated errors and lagged dependent variables in its regressor set, the OLS estimators for H_1 and H_2 are, in general, both inconsistent. The limited but important purpose of this paper is to help applied workers to detect situations in which OLS estimators cannot be assumed to be consistent and asymptotically normal, with heteroskedasticity-robust inference being available. Hence, in the experiments of this section, the desired outcome of the test procedure is rejection of the intersection null hypothesis that H_1 is the correct model and has independent errors. (The objective of providing tests of H_1 against H_2 that are robust against autocorrelation is much more ambitious and would seem to require OLS to be abandoned in favour of IV, as discussed in Section 3.1 above.)

The data are generated by a DGP that consists of (14) with $k_2 = 2$ and errors given by $u_{2t} = |z_{2t}|v_t$, in which the variate v_t is obtained from the autoregression (16). A sample of results is provided in Table 6, with $T = 40$, and (16) having the two sets of values of ϕ_j, $j = 1, ..., 4$, given in Table 5. (The error models in Tables 5 and 6 do not use the same value of σ_a^2 because rejection rates are too close to 100 per cent to provide interesting comparisons when $\sigma_a^2 = 0.1$ is used with H_2.)

TABLE 6

Rejection frequencies, with $T = 40$ and a nominal significance level of 5 per cent, when the DGP is H_2 with autocorrelated and heteroskedastic errors

(a) Coefficients of (16): $\phi_1 = 0.75$; $\phi_2 = 0$; $\phi_3 = 0$; $\phi_4 = 0$; $\sigma_a^2 = 1.0$

				$m = 1$		$m = 4$	
ψ_2	k_1	k_2	Value of ρ^2 is	0.3	0.9	0.3	0.9
0.3	3	2		63.0	31.0	42.6	25.1
0.5	3	2		64.2	34.2	42.8	24.7
0.7	3	2		67.6	42.8	45.8	30.0

(b) Coefficients of (16): $\phi_1 = 0.7$; $\phi_2 = -0.17$; $\phi_3 = 0.017$; $\phi_4 = -0.0006$; $\sigma_a^2 = 1.0$

				$m = 1$		$m = 4$	
ψ_2	k_1	k_2	Value of ρ^2 is	0.3	0.9	0.3	0.9
0.3	3	2		72.6	41.0	49.9	28.0
0.5	3	2		74.0	44.7	51.2	30.1
0.7	3	2		76.8	50.8	53.8	34.2

Notes: The test of the untrue intersection null hypothesis is carried out using the recursive wild bootstrap method, with the Rademacher pick distribution. All results are reported as percentages, rounded to one decimal place.

As in Table 5, every rejection frequency for the joint test with $m = 1$ in Table 6 is greater than the corresponding value that is obtained with $m = 4$. As expected, the rejection frequencies decrease as ρ^2 (the squared population correlation coefficient between x_{2t} and z_{2t}) increases, when other design parameters are held constant. Also, in most cases, rejection rates increase with ψ_2, other things being equal, but not by large amounts.

6 Conclusions

The problem of testing nonnested models in which the regressors include lagged dependent variables has been discussed under the assumption that estimates and tests are derived using OLS. The application of OLS techniques to models with some lagged dependent variables as regressors is common in applied work. However, standard results concerning the consistency of OLS estimators and the asymptotic validity of associated confidence intervals and tests require that the errors are not autocorrelated. It follows that OLS results cannot be used to obtain heteroskedasticity and autocorrelation consistent tests of nonnested dynamic regression models. However, if the errors are independent, it is possible to derive OLS-based tests that are asymptotically valid in the presence of unspecified forms of conditional heteroskedasticity. It would clearly be wrong to assume independence without examining the strength of the evidence that the sample provides against this assumption, given that the consequences of autocorrelation are so serious in models of the type discussed in this paper.

It has, therefore, been argued that, when lagged dependent variables are regressors, it is essential to check for autocorrelation of the errors, as well as to test the specification of the regression function using information about the nonnested alternative model. An approach to implementing a heteroskedasticity-robust joint test has been proposed. Monte Carlo evidence suggests that a recursive wild bootstrap method produces good control of the finite sample significance levels of the heteroskedasticity-robust joint test, but asymptotic critical values fail to give reliable approximations. The general strategy of using a wild bootstrap with heteroskedasticity-robust tests to detect misspecifications that lead to the inconsistency of OLS estimators could be applied in many other contexts. The key requirement is that the diagnostic checks that are to be combined should be capable of being calculated by testing the significance of variables that are added to the regressors of the model under scrutiny.

References

Andrews, D. W. K. (1991). 'Heteroskedasticity and autocorrelation consistent covariance matrix estimation', *Econometrica*, Vol. 59, pp. 817-858.

Bose, A (1988). 'Edgeworth correction by bootstrap in autoregressions', *Annals of Statistics*, Vol. 16, pp. 1709-1722.

Breusch, T. S. (1978). 'Testing for autocorrelation in dynamic linear models', *Australian Economic Papers*, Vol. 17, pp. 334-355.

Choi, H.-S., and Kiefer, N. M. (2008). 'Robust nonnested testing and the demand for money', *Journal of Economic & Business Statistics*, Vol. 26, pp. 9-17.

Cribari-Neto, F. (2004). 'Asymptotic inference under heteroskedasticity of unknown form,' *Computational Statistics and Data Analysis*, Vol. 45, pp. 215-233.

Darroch, J. N. and Silvey, S. D. (1963). 'On testing more than one hypothesis,' *Annals of Mathematical Statistics*, Vol. 34, pp. 555-567.

Davidson, J., Monticini, A. and Peel, D. (2007). 'Implementing the wild bootstrap using a two-point distribution,' *Economics Letters*, Vol. 96, pp. 309-315.

Davidson, R. and Flachaire, E. (2008). 'The wild bootstrap, tamed at last,' *Journal of*

Econometrics, Vol. 146, pp. 162-169.

Davidson, R., and MacKinnon, J. G. (1981). 'Several tests for model specification in the presence of alternative hypotheses,' *Econometrica*, Vol. 49, pp. 781-793.

Davidson, R. and MacKinnon, J. G. (1985a). 'Heteroskedasticity-robust tests in regression directions,' *Annales de l'INSEE*, Vol. 59/60, pp. 183-218.

Davidson, R. and MacKinnon, J. G. (1985b). 'The interpretation of test statistics,' *Canadian Journal of Economics*, Vol. 18, pp. 38-57.

Delgado, M. A. and Stengos, T. (1994). 'Semiparametric specification testing of non-nested econometric models,' *Review of Economic Studies*, Vol. 61, pp. 291-303.

Fan, Y. and Li, Q. (1995). 'Bootstrapping J-type tests for non-nested regression models,' *Economics Letters*, Vol. 48, pp. 107-112.

Flachaire, E. (2005). 'Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap,' *Computational Statistics and Data Analysis*, Vol. 49, pp. 361-376.

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

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

Godfrey, L. G. (2007). 'Alternative approaches to implementing Lagrange multiplier tests for serial correlation in dynamic regression models,' Computational Statistics and Data Analysis, Vol. 51, pp. 3282-3295.

Godfrey, L. G. and Orme, C. D. (2004). 'Controlling the finite sample significance levels of heteroskedasticity-robust tests of several linear restrictions on regression coefficients,' *Economics Letters*, Vol. 82, pp. 281-287.

Godfrey, L. G. and Pesaran, M. H. (1983). 'Tests of non-nested regression models: small sample adjustments and Monte Carlo evidence,' *Journal of Econometrics*, Vol. 21, pp. 133-154.

Godfrey, L. G. and Tremayne, A. R. (1988). 'Checks of model adequacy for univariate time series models and their application to econometric relationships,' Econometric Reviews, Vol. 7, pp. 1-42.

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

Gonçalves, S. and Kilian, L. (2004). 'Bootstrapping autoregressions with conditional heteroskedasticity of unknown form,' *Journal of Econometrics*, Vol. 123, pp. 89-120.

Greene, W. H. (2008). *Econometric Analysis*, 6th Edition. New Jersey: Prentice-Hall.

Hansen, B. E. (1999). 'Discussion of "Data mining reconsidered",' *Econometrics Journal*, Vol. 2, pp. 192-201.

Horowitz, J. L. and Savin, N. E. (2000). 'Empirically relevant critical values for hypothesis tests: A bootstrap approach,' *Journal of Econometrics*, Vol. 95, pp. 375-389.

Hsieh, D. S. (1983). 'A heteroscedasticity-consistent covariance matrix estimator for time series regressions,' *Journal of Econometrics*, Vol. 22, pp. 281-290.

Kiefer, N. M., Vogelsang, T. J. and Bunzel, H. (2000). 'Simple robust testing of regression hypotheses,' *Econometrica*, Vol. 68, pp. 695-714.

Kiefer, N. M., and Vogelsang, T. J. (2002a). 'Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation,' *Econometrica*, Vol. 70, pp. 2093-2095.

Kiefer, N. M., and Vogelsang, T. J. (2002b). 'Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size,' *Econometric Theory*, Vol. 18, pp. 1350-1366.

Kiefer, N. M., and Vogelsang, T. J. (2005). 'A new asymptotic theory for heteroskedasticity-autocorrelation robust tests,' *Econometric Theory*, Vol. 21, pp. 1130-1164.

Liu, R. Y. (1988). 'Bootstrap procedures under some non i.i.d. models,' *Annals of Statistics*, Vol. 16, pp. 1696-1708.

Long, J. S. and Ervin, L. H. (2000). 'Using heteroskedasticity-consistent standard errors in the linear regression model,' *American Statistician*, Vol. 32, pp. 217-224.

McAleer, M. (1995). 'The significance of testing empirical non-nested models,' *Journal of Econometrics*, Vol. 67, pp. 149-171.

MacKinnon, J. G. and White, H. (1985). 'Some heteroscedasticity consistent covariance matrix estimators with improved finite sample properties,' *Journal of Econometrics*, Vol. 29, pp. 305-325.

MacKinnon, J. G., White, H., and Davidson, R. (1983). 'Tests for model specification in the presence of alternative hypotheses: some further results,' *Journal of Econometrics*, Vol. 21, pp. 53-70.

Mammen, E. (1993). 'Bootstrap and wild bootstrap for high dimensional linear models,' *Annals of Statistics*, Vol. 21, pp. 255-285.

Mizon, G. E. (1995). 'A simple message for autocorrelation correctors: Don't,' *Journal of Econometrics*, Vol. 69, pp. 267-288.

Mizon, G. E. and Richard, J.-F. (1986). 'The encompassing principle and its application to testing non-nested hypotheses,' *Econometrica*, Vol. 54, pp. 657-678.

Pesaran, M. H. and Weeks, M. (2001). 'Non-nested hypothesis testing: an overview,' in *Companion to Theoretical Econometrics*, ed. B. H. Baltagi, Oxford: Basil Blackwell, pp. 279-309.

Serlin, R. C. (2000). 'Testing for robustness in Monte Carlo studies,' *Psychological Methods*, Vol. 5, pp. 230-240.

Stock, J. H., and Watson, M. W. (2006). *Introduction to Econometrics*, 2nd edition, Boston: Pearson.

White, H. (1980). 'A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity,' *Econometrica*, Vol. 48, pp. 817-838.