Empirical Economics 31:631–643 (2006)
DOI 10.1007/s00181-005-0039-0

ORIGINAL PAPER

Dimitris Hatzinikolaou · Athanassios Stavrakoudis

Empirical size and power of some diagnostic tests applied to a distributed lag model

Received: 25 November 2004 / Accepted: 12 July 2005 / Published online: 5 April 2006

Abstract We produce Monte Carlo evidence on the size and power of the RESET, a heteroscedasticity test, and a test for autocorrelation applied to realistic distributed-lag models. We find that the autocorrelation test has the correct size and high power to detect not only autocorrelation (given a correct model), but also the erroneous omission of *several* lags of an explanatory variable, whereas the RESET and heteroscedasticity tests are oversized in the presence of positive disturbance autocorrelation, especially when the regressors are also positively autocorrelated, and have no power to detect such misspecification errors. In large samples, size distortion may be avoided by using autocorrelation-robust methods.

Keywords Size · Power · Simulation · RESET · Diagnostic tests

JEL Classification C15 · C22 · C52

1 Introduction

The empirical results of a particular econometric application are often deemed credible if a small number of well known diagnostic tests fail to provide strong evidence against the underlying assumptions of the estimated model. Such a strategy is not always acceptable, however, because these tests do not have high power against every alternative. Another important problem is that these tests may suffer from size distortion when a classical assumption is violated. This problem is closely related to that of correct assessment of power, since power estimates are correct only if the actual and the nominal size are approximately equal. For

We gratefully acknowledge the constructive comments of an anonymous referee of this journal, which improved significantly the paper. We also thank our colleagues S. Symeonides and E. Zacharias for their comments. The usual disclaimer applies.

D. Hatzinikolaou (✉) · A. Stavrakoudis
Department of Economics, University of Ioannina, 451 10, Ioannina, Greece
E-mail: dhatzini@cc.uoi.gr

example, the power of an oversized test may be overstated; see Kiviet (1986, p. 254) and Godfrey et al. (1988, pp. 497–499).

The size and power of diagnostic tests, especially Ramsey's (1969) regression specification error test (RESET), have been extensively investigated by Monte Carlo methods; see, e.g., Thursby and Schmidt (1977), Thursby (1979, 1989), Porter and Kashyap (1984), Kiviet (1986), Godfrey et al. (1988), Godfrey and Orme (1994), and Leung and Yu (2001). To our knowledge, the literature has so far considered only simple dynamic models, however, and neglected more general autoregressive distributed lag models, where, for example, several lagged values of an explanatory variable are erroneously omitted from the null model. This is surprising, since we often encounter such models in empirical economics, e.g., trade balance equations incorporating the J-curve effect, inflation equations, etc.; and since the erroneous omission of dynamics, e.g., in the spirit of the general-to-specific approach, will in general lead to invalid statistical inference (Kiviet 1986, p. 243). In addition, papers that consider simple dynamic models, e.g., Kiviet (1986) and Thursby (1989), run only a small number of replications (500 and 200, respectively), and do not study the behavior of the variants of the RESET proposed by Ramsey (1969) and used most frequently in applied work. In fact, Kiviet (1986) excludes altogether the RESET from his list of diagnostic tests.

In this paper, we consider more realistic distributed lag models with two explanatory variables, possibly collinear and autocorrelated, and possibly autocorrelated disturbances. We use 5,000 replications to investigate the effects of changing the parameters of the data generating process on the size and power of five diagnostic tests: three variants of Ramsey's RESET, a test for autocorrelation, and a test for heteroscedasticity. We use both ordinary least squares (OLS) and autocorrelation-robust methods, since the RESET is known to be sensitive to the presence of autocorrelation; see Porter and Kashyap (1984) and Leung and Yu (2001).

We find that the test for autocorrelation has approximately the correct size and high power to detect not only autocorrelation (given a correct model), but also the erroneous omission of several lags of an explanatory variable. In contrast, the RESET and heteroscedasticity tests are oversized in the presence of positive error autocorrelation, especially when the regressors are also positively autocorrelated, and have no power to detect such misspecification errors. When we use autocorrelation-robust methods, the size distortion of the RESET is reduced drastically, but its power deteriorates even further. After describing the five tests (Section 2) and our Monte Carlo setup (Section 3), we report our results and offer some possible explanations for the reported patterns (Section 4). Section 5 concludes the paper.

2 The diagnostic tests

Consider the linear regression model $\mathbf{y} = \mathbf{X}\beta + \mathbf{v}$, where $\mathbf{y}$ and $\mathbf{v}$ are $T\times1$ vectors, β is a $K\times1$ vector of coefficients, and $\mathbf{X}$ is a $T\times K$ matrix of rank K containing T observations on K regressors. The disturbances, ν_t, $t = 1, \ldots, T$, are assumed to be normally distributed; and, under the null hypothesis of correct specification, they are serially independent with zero mean and constant variance, conditional on $\mathbf{X}$. We consider only stationary time-series data.

This is the null model considered for regression, assuming that the OLS assumptions hold. We consider the following diagnostic tests, which are used widely in the literature for checking the adequacy of OLS regression equations. In each experiment, if p-value < 0.05, we record one rejection of the corresponding H_0, i.e., we use a 5% nominal level of significance.

The RESET This test is intended to detect any specification errors that may be present in a linear null model, e.g., omitted variables, incorrect functional form, and correlation between $\mathbf{X}$ and $\mathbf{v}$. If one or more of these errors are present, the main OLS assumption, $E(\mathbf{v}|\mathbf{X}) = 0$, is violated, i.e., $E(\mathbf{v}|\mathbf{X}) = \xi \neq 0$, where ξ denotes the omitted portion of the regression. The consequences are serious (Krämer et al. 1985). To test the null hypothesis H_0: $E(\mathbf{v}|\mathbf{X}) = 0$, choose a $T \times M$ matrix $\mathbf{Z}$ of "test variables," apply OLS to the equation

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{u}, \tag{1}$$

and test the hypothesis H_0: $\gamma = 0$ using a standard F test.

The test variables in $\mathbf{Z}$ are usually chosen following Ramsey's (1969) suggestion, which we also adopt in this paper: $z_t = \widehat{Y}_t^2$, or $\mathbf{z}_t = \left(\widehat{Y}_t^2, \widehat{Y}_t^3\right)$, or $\mathbf{z}_t = \left(\widehat{Y}_t^2, \widehat{Y}_t^3, \widehat{Y}_t^4\right)$, where $\widehat{Y}_t = \mathbf{x}_t' \widehat{\beta}$ is the OLS predicted value of Y_t obtained from the null model, and $\mathbf{x}_t$ and $\mathbf{z}_t$ are the t-th rows of the matrices $\mathbf{X}$ and $\mathbf{Z}$. Since these test variables are the second, third, and fourth powers of $\hat{Y}_t$, we denote these variants of RESET as POY2, POY3, and POY4, respectively. Note that the variant POY2 is recommended by Godfrey et al. (1988, pp. 501–502) and Godfrey and Orme (1994, p. 506) and is used by the popular econometric program Microfit.

Thursby and Schmidt (1977, p. 635) derive two conditions for the RESET to have some power under the alternative hypothesis. If $\xi \neq 0$ then $E(\widehat{\gamma}|\mathbf{X}) = (\mathbf{Z}'\mathbf{M}_x\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_x\xi$, where $\mathbf{M}_x = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. But $E(\widehat{\gamma}|\mathbf{X})$ will be non-zero if $\mathbf{Z}'\mathbf{M}_x\xi \neq 0$, which can hold if either of the following two conditions holds: (1) $\mathbf{Z}$ and ξ are correlated, in that $\mathbf{Z}'\xi \neq 0$ or (2) even if $\mathbf{Z}'\xi = 0$, both $\mathbf{Z}$ and ξ are correlated with $\mathbf{X}$. Thursby and Schmidt (1977, p. 637) note, however, that the power of the test generally decreases as the correlation between ξ and $\mathbf{X}$ increases. For since $\mathbf{X}$ is included in the regression, the higher is its correlation with the omitted variables, ξ, the less information is lost when the variables ξ are omitted from the regression, and the more difficult it becomes for the test to detect this loss.

If the null model erroneously omits nonlinear terms involving the included variables, then the RESET has high power, because in this case the test variables are approximate functions of the omitted nonlinearities, so $\mathbf{Z}'\xi \neq 0$, and condition (1) holds; see Thursby and Schmidt (1977, p. 639) and Thursby (1979, p. 224, and 1989, pp. 227–229). But for other types of model inadequacy, e.g., omitted variables or omitted lags, the above two conditions might not hold, and the RESET may have low power; see Thursby's (1989) Tables 5, 7, and 8.

The test for autocorrelation This test, denoted here as LMF, is the F-version of the well known Lagrange Multiplier test for autocorrelation (Wooldridge 2003, p. 399). We use it to test the hypothesis H_0: "there is no autocorrelation in the disturbances" against the alternative H_1: "there is first-order autocorrelation." The

test is carried out in three steps. First, apply OLS to the null model and obtain the residuals, $\widehat{v}_t$. Second, apply OLS to the equation

$$\widehat{v}_t = \mathbf{x}_t'\alpha + \gamma\widehat{v}_{t-1} + error, \tag{2}$$

where $\mathbf{x}_t$ is the same $K\times1$ vector of the regressors used in the null model (including a constant term), and α is a $K\times1$ coefficient vector. Third, test the hypothesis H_0: $\gamma=0$ against H_1: $\gamma\neq0$.

This test is sometimes viewed as a general misspecification test, because it has power in detecting omitted variables and incorrect functional form; see Thursby (1979, p. 222) and Kiviet (1986, pp. 254–255), who considers n-th order autocorrelation, n=1, 4, 8. Note Kiviet's finding that the LMF test has no power to detect the omission of x_t and x_{t-1} from the true model $y_t = \gamma y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t$ (1986, last two columns of Table VII, pp. 254–255).

The test for heteroscedasticity This test, denoted here as HET, is designed to test the hypothesis H_0: "the disturbances are homoscedastic" against the alternative H_1: $Var(v_t|\mathbf{X}) = [E(Y_t|\mathbf{X})]^2$. To carry out the test, first apply OLS to the null model and obtain the residuals and the predicted values, $\widehat{v}_t$ and $\widehat{Y}_t$, as well as their squares. Then, apply OLS to the equation

$$\widehat{v}_t^2 = \gamma_0 + \gamma_1\widehat{Y}_t^2 + error, \tag{3}$$

and test the hypothesis H_0: $\gamma_1 = 0$ against H_1: $\gamma_1 \neq 0$. This test is also sometimes viewed as a general misspecification test, since misspecification can cause evidence of heteroscedasticity; see Pagan and Hall (1983, pp. 168, 178) and Wooldridge (2003, pp. 269–270).

3 Monte Carlo design

We use the following true model in our Monte Carlo experiments:

$$Y_t = 0.0 - X_{1t} + 0.8X_{2t} + 0.6X_{1t-1} + 0.5X_{1t-2} + 0.4X_{1t-3} + 0.3X_{1t-4} + u_t, \tag{4}$$

where the data for X_{1t}, X_{2t} and u_t are generated as follows:

$$X_{1t} = \varphi X_{1t-1} + \varepsilon_{1t}, \tag{5}$$

$$X_{2t} = \theta X_{1t} + \varepsilon_{2t}, \tag{6}$$

$$u_t = \rho u_{t-1} + w_t, \tag{7}$$

$$\varepsilon_{1t}, \varepsilon_{2t} \sim i.i.d.N(5, 10), w_t \sim i.i.d.N(0, 1), \tag{8}$$

$$X_{10} \sim N\big(5/(1-\varphi), 10/\big(1-\varphi^2\big)\big),\ u_0 \sim N\big(0, 1/\big(1-\rho^2\big)\big), \tag{9}$$

$$\varphi = 0.0, 0.5, 0.95,\ \theta = 0.0, 0.2, 0.9,\ \rho = 0.0, 0.5, 0.95. \tag{10}$$

Eq. (4) can be thought of as a trade balance equation capturing the J-curve effect, where Y = net exports, X_1 = real exchange rate (the ratio of prices of foreign goods, measured in domestic currency, relative to prices of home goods), and X_2 = the ratio of foreign to domestic GDP. According to Eq. (4), a real depreciation of the domestic currency (an increase in X_{1t}), which makes foreign goods more expensive relative to home goods, initially worsens the home country's trade balance, but the total effect four periods later is positive.

For each sample size, T=50 and T=200, and each combination of the parameters φ, θ, and ρ, we generate *one* set of T "observations" for each of the ε_1's and ε_2's and 5,000 sets for the w's, all from normal distributions, as indicated.[1] Then, we construct one set of "observations" for X_{1t} and X_{2t}, which we keep fixed in the 5,000 replications; use the true model to generate 5,000 sets of T "observations" for the dependent variable; and use these "observations" to estimate 5,000 times the null model,

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \sum\nolimits_{j=1}^{4} \beta_{j+2} X_{1t-j} + \nu_t. \tag{11}$$

We estimate the power of a diagnostic test as the proportion of rejections obtained when the test is applied 5,000 times to Eq. (11), after placing one of the following four sets of restrictions: (1) $\beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$, (2) $\beta_4 = \beta_5 = \beta_6 = 0$, (3) $\beta_5 = \beta_6 = 0$, and (4) $\beta_6 = 0$. If no restrictions are placed on the β's, however, Eq. (11) is the correct model, so this proportion of rejections estimates the true size of the test, provided that there is no autocorrelation in the disturbances. (It would also estimate size if, instead of the correct model, we estimated an over-parameterized version of it; see Kiviet 1986, p. 249.) In the presence of autocorrelation, however, although this proportion clearly estimates power when it refers to an autocorrelation test, its meaning is less clear when it refers to other diagnostic tests, e.g., the RESET and HET, since these two tests are usually viewed as tests against specific alternatives rather than as general misspecification tests.

For example, the RESET is usually viewed as a test of the hypothesis $E(\mathbf{v}|\mathbf{X}) = 0$, which can hold despite the presence of autocorrelation. Thus, in Monte Carlo studies applying the RESET to an equation that has autocorrelated disturbances, but is otherwise correctly specified, if the rejection frequency is higher than the nominal size, the RESET is said to lack robustness to autocorrelation and to be an oversized test. High rejection frequencies of the RESET can occur because the conventional t and F tests of significance are not robust to positive disturbance autocorrelation and tend to over-reject, especially when the regressors are also positively autocorrelated; see Johnston (1972, pp. 248–249) and Krämer et al. (1990).

The RESET may also lack robustness because of the well known problem of "spurious correlation," which can arise if a regressor X_{jt} and the error term u_t are both highly autocorrelated (Leung and Yu 2001). But failure of the assumption $Corr(X_{jt}, u_t) = 0$ implies failure of H_0: $E(\mathbf{u}|\mathbf{X}) = 0$ (Wooldridge 2003, p. 719); and since autocorrelation ($\rho > 0$) is partly responsible for this failure, it follows that, when applied to the correct model, the RESET is likely to reject more

[1] We use the LINUX version of RATS v. 5.01 to carry out the simulations. Random numbers were generated using the function %RAN(x) and the starting seed 317811. Note also that before we started drawing the values of ε_1, ε_2, and w, we let the process run for 500 "periods."

frequently as ρ takes on higher values. To see why, note that since the test variables **Z** depend on **X**, it follows that if **X** and **u** are (spuriously) correlated, so that $\xi = E(\mathbf{u}|\mathbf{X})$ depends on **X**, then **Z** and ξ will also be correlated. Thus, condition (1) of the previous section is satisfied, and the RESET may over-reject the correct model. Since we have "observations" on the error term u_t, we can calculate the correlation coefficients $Corr(X_{jt}, u_t)$ and thus assess the severity of the "spurious correlation" problem.

This problem may also cause HET to over-reject the correct model. For if a regressor X_{jt} is (spuriously) correlated with the error term, u_t, it may also be correlated with u_t^2 . And since $\widehat{Y}_t$ is correlated with X_{jt}, it is possible that $\widehat{Y}_t^2$ might have explanatory power in Eq. (3).

To be sure, tests designed against specific alternatives, also called "one-directional tests," assume that all the other standard assumptions are satisfied and, strictly speaking, are not valid in the presence of other misspecifications (Godfrey 1988, p. 107). For example, the HET test is not valid unless $\rho = 0$ (Wooldridge 2003, pp. 414–415). Its application despite the presence of autocorrelation can only be justified if it is viewed not as a "one-directional test," but as a general misspecification test, as was noted in the previous section.

Nevertheless, we will interpret the rejection frequencies of the RESET and HET tests applied to the correct model with $\rho > 0$ as empirical size, saving the term "power" for their application to an incorrect model. And if the empirical size of these tests is significantly greater than the nominal size when $\rho > 0$, we will say that the tests are "sensitive" to autocorrelation. This sensitivity is desirable in that the tests can thus be viewed as general misspecification tests, but undesirable in that it becomes difficult to separate misspecification from autocorrelation (Leung and Yu 2001, p. 726).

In applied work, it is important to remember that some diagnostic tests may be sensitive to misspecifications for which they were not designed (Godfrey 1988, p. 4). Thus, when a diagnostic test rejects, the model must be re-specified, although not necessarily in accordance with the alternative that the test is designed to detect. For example, "a significant value of a statistic testing for serial independence need not imply that the model should be augmented by a serially correlated error process" (Kiviet 1986, p. 243). It may imply that an autocorrelated variable, or a lagged value of it, has been erroneously omitted from the null model.

Finally, note that since we use a 5% nominal level of significance and 5,000 replications, the standard 95% confidence interval for the true percentage of rejections of a diagnostic test is (4.40, 5.60) (Godfrey and Orme 1994, p. 498). Estimated sizes that fall outside this interval are regarded as significantly different from the nominal size.

4 Results

Estimation by OLS Table 1 reports rejection frequencies of the five diagnostic tests (POY2, POY3, POY4, LMF, and HET) generated from Eq. (11) by OLS with $T = 50$. The rejection frequencies reported under "correct model" estimate test size and emerge when no restrictions are imposed, whereas those under "misspecified null

Table 1 Rejection frequencies of the five diagnostic tests generated by applying OLS to Eq. (11) with $T = 50$

			Correct model (unrestricted Eq. (11))					Misspecified null model (restrictions: $\beta_3=\beta_4=\beta_5=\beta_6=0$)				
φ	θ	ρ	POY2	POY3	POY4	LMF	HET	POY2	POY3	POY4	LMF	HET
0.00	0.00	0.00	5.10	4.86	4.68	4.30*	4.66	0.00	0.40	0.02	100.00	0.08
		0.50	6.70*	5.84*	5.30	80.64	2.88*	0.00	2.36	0.52	100.00	0.06
		0.95	7.24*	3.70*	3.48*	99.92	0.50*	0.00	20.62	16.94	100.00	0.36
	0.20	0.00	4.92	5.62*	4.90	4.38*	5.42	0.28	0.16	0.00	100.00	0.08
		0.50	3.54*	2.30*	2.82*	81.66	4.00*	0.46	0.14	0.00	100.00	0.02
		0.95	0.98*	0.30*	0.26*	99.98	1.50*	3.34	0.42	0.02	100.00	0.12
	0.90	0.00	4.98	4.88	5.14	4.58	5.04	14.78	6.84	13.76	100.00	0.48
		0.50	4.46	4.96	4.50	82.26	5.62*	16.44	8.28	17.16	100.00	0.68
		0.95	2.42*	3.52*	2.38*	99.92	4.76	24.26	14.74	11.20	100.00	4.76
0.50	0.00	0.00	5.26	5.22	4.84	4.48	4.68	0.00	0.00	0.00	100.00	0.00
		0.50	4.92	3.98*	3.72*	75.62	5.46	0.00	0.00	0.00	100.00	0.00
		0.95	7.60*	5.26	3.60*	99.92	6.66*	0.02	0.00	0.06	100.00	1.20
	0.20	0.00	5.20	5.14	5.02	4.14*	3.98*	0.04	0.00	0.02	100.00	0.00
		0.50	5.96*	4.48	5.46	74.46	5.46	0.34	0.00	0.06	100.00	0.00
		0.95	14.20*	5.44	5.04	99.88	10.46*	1.66	0.06	2.20	100.00	1.06
	0.90	0.00	4.78	4.88	4.96	4.82	4.92	0.00	0.00	0.00	100.00	0.00
		0.50	9.42*	11.34*	9.16*	77.32	6.84*	0.00	0.00	0.00	100.00	0.00
		0.95	21.20*	28.22*	20.16*	99.94	16.80*	0.00	0.00	0.00	100.00	0.02
0.95	0.00	0.00	4.96	4.92	5.20	4.80	4.08*	0.00	82.86	65.32	100.00	0.00
		0.50	13.38*	11.84*	11.70*	73.10	6.66*	0.00	77.46	62.90	100.00	0.00
		0.95	26.74*	21.26*	22.40*	99.84	10.60*	0.02	71.84	57.36	100.00	0.10
	0.20	0.00	5.08	4.72	4.78	4.82	4.46	0.00	34.04	1.70	100.00	0.00
		0.50	14.54*	14.20*	12.76*	74.90	5.62*	0.00	37.70	3.46	100.00	0.02
		0.95	31.64*	28.68*	26.06*	99.76	10.58*	0.00	43.74	20.60	100.00	0.20
	0.90	0.00	5.14	4.86	4.80	5.14	3.62*	0.00	51.24	1.84	100.00	0.00
		0.50	20.52*	27.16*	28.02*	73.38	6.62*	0.00	52.88	4.82	100.00	0.00
		0.95	56.56*	67.02*	68.90*	99.68	11.12*	0.02	53.84	17.20	100.00	0.04

(a) The rejection frequencies are given in percentages; (b) the frequencies that fall outside the interval (4.40, 5.60) are regarded as significantly different from the nominal size (5%) and are marked by a star (*); (c) the entries for LMF under the heading "correct model" represent size when ρ=0 and power against autocorrelation when ρ>0; (d) the entries under the heading "misspecified null model" represent power against the alternative of erroneously imposing the restrictions $\beta_3=\beta_4=\beta_5=\beta_6=0$.

model" estimate test power and emerge when the following four restrictions are imposed: $\beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$.

Consider size. When $\rho = 0$, the estimated size of all the five tests ranges from 3.62 to 5.62%, with only seven significant departures from the nominal size in 45 experiments. Thus, in the absence of autocorrelation, the five tests have approximately the correct size, regardless of the values of φ and θ. But when $\rho = 0.5$ or $\rho = 0.95$, the estimated size of the RESET ranges from 0.26 to 68.90%, with the most extreme estimates occurring at $\rho = 0.95$, and increases as φ increases.

Thus, as was expected from our earlier discussion, when the regressors and the error term are both positively autocorrelated, the RESET is oversized.

The value of θ also matters. When $\varphi = 0$, the actual size of RESET falls significantly below the nominal size at $\theta = 0.2$, but when $\varphi = 0.5$ or $\varphi = 0.95$, this size increases as θ increases, reaching its highest values when $\varphi = 0.95$, $\rho = 0.95$, and $\theta = 0.9$. Higher values of θ allow more of the autocorrelation in X_1 to be passed on to X_2, since Eq. (6) implies that $\rho_{12} = \theta\sigma_1/\sigma_2$, where ρ_{12} is the correlation coefficient between the variables X_1 and X_2, and σ_1 and σ_2 are their standard deviations. Thus, when $\varphi = 0.95$ and $\rho = 0.95$, the higher is the value of θ the more severe becomes the problem caused by simultaneous positive autocorrelation in the regressors and in the disturbances. For the same reason, the "spurious correlation" problem discussed earlier may also become more severe. For if $Corr(X_{1t}, u_t)$ is high at $\varphi = 0.95$ and $\rho = 0.95$, $Corr(X_{2t}, u_t)$ is also expected to be high at high values of θ. To assess the severity of this problem, we calculate the frequency distributions of $Corr(X_{jt}, u_t), j = 1, 2$, at the various combinations of φ, ρ, and θ. Using the standard test of significance of a correlation coefficient (Johnston 1972, pp. 36–37), we find, for example, that when $\varphi = 0.95$, $\rho = 0.95$, $\theta = 0.2$, and $T = 50$, more than 58% of the values of $Corr(X_{1t}, u_t)$ are statistically different from zero. Thus, the "spurious correlation" problem may explain, at least in part, why the RESET is oversized.

In the presence of autocorrelation, the HET test also suffers from size distortion, but the problem is less severe here. The actual size of HET ranges from 0.5 to 16.8%, and is higher at $\rho = 0.95$ than at $\rho = 0.5$ only as long as $\varphi = 0.5$ or $\varphi = 0.95$. This pattern can also be explained by invoking the non-robustness of the t and F tests to autocorrelation. In particular, we find that when $\varphi = 0.5$ or $\varphi = 0.95$, both the explanatory variable and the residuals in Eq. (3) are positively autocorrelated, especially when $\rho = 0.95$.

Next, consider power. The power of RESET to detect the erroneous omission of the four lags of X_{1t} is disappointing: POY2 has low or zero power at every parameter combination, whereas POY3 and POY4 have some power at $\varphi = 0.95$. Thus, in the present setup, the claim of Godfrey et al. (1988, p. 501) and Godfrey and Orme (1994, pp. 500–501), that POY2 has better power than POY3 and POY4, does not hold true.

Having in mind conditions (1) and (2) for the RESET to have some power (see Section 2), we have estimated the following three types of regressions (all of which include an intercept): (a) $\widehat{Y}_t^j$ on $X_{1t-1}, X_{1t-2}, X_{1t-3}$ and X_{1t-4}, where $\widehat{Y}_t^j$ is the j-th power of $\widehat{Y}_t$, $j = 2, 3, 4$; (b) $\widehat{Y}_t^j$ on X_{1t} and $X_{2t}, j = 2, 3, 4$; and (c) X_{1t-j} on X_{1t} and $X_{2t}, j = 1, 2, 3, 4$. Let the R^2's from these regressions be denoted as $R^2_{\mathbf{Z}.\xi}$, $R^2_{\mathbf{Z}.\mathbf{X}}$, and $R^2_{\xi.\mathbf{X}}$. (Note that the values of $R^2_{\mathbf{Z}.\xi}$ and $R^2_{\mathbf{Z}.\mathbf{X}}$ are averages from 5,000 replications, but there is just one value for $R^2_{\xi.\mathbf{X}}$, since the same set of X's is used in every replication.) Now the above two conditions can be restated as follows: (1) $R^2_{\mathbf{Z}.\xi} > 0$; or (2) even if $R^2_{\mathbf{Z}.\xi} = 0$, both $R^2_{\mathbf{Z}.\mathbf{X}} > 0$ and $R^2_{\xi.\mathbf{X}} > 0$.

We find that for $\varphi = 0$ or $\varphi = 0.5$, the values of $R^2_{\mathbf{Z}.\xi}$ and $R^2_{\xi.\mathbf{X}}$ are low, ranging from 0.03 to 0.22; whereas for $\varphi = 0.95$, they are high, ranging from 0.73 to 0.86. These values are consistent with the fact that the RESET has some power only when $\varphi = 0.95$, despite the fact that the values of $R^2_{\mathbf{Z}.\mathbf{X}}$ are high at every parameter combination, ranging from 0.65 to 0.99.

Now consider the power of HET. As the last column of Table 1 shows, this is a biased test, since its power is less than its size at every parameter combination, except at ($\varphi = 0$, $\rho = 0.95$, $\theta = 0.9$), where size and power are both 4.76%.

Finally, consider the power of LMF. The rejection frequencies under "correct model" and for $\rho = 0.5$ or $\rho = 0.95$ estimate its power against autocorrelation, whereas those under "misspecified null model" estimate its power against misspecification, especially at $\rho = 0$. Both powers are high. When $\rho = 0.5$, the power of LMF against autocorrelation ranges from 73.10 to 82.26%, and when $\rho = 0.95$ it is at least 99.68%. Against misspecification, LMF has 100% power at every parameter combination.

For space considerations, we do not report the results for $T = 200$.[2] The main differences from the case of $T = 50$ are as follows. First, when $\rho = 0$, *all* size estimates now fall in the interval (4.4, 5.6). Thus, when there is no disturbance autocorrelation, the five tests are exact in large samples, i.e., they have the correct size, regardless of the values of φ and θ.

Second, when $\rho = 0.5$ or $\rho = 0.95$, the size of the RESET and HET tests is now generally higher than in the case of $T = 50$, ranging from 1 to 75%. The highest size estimates occur when $\varphi = 0.95$, $\rho = 0.95$, and $\theta = 0.90$. Porter and Kashyap (1984, pp. 231–232) also find that the size distortion of the RESET can increase as the sample size increases.

Third, the only case now where the RESET has some power (ranging from 34 to 89%) is when $\varphi = 0.5$ and $\theta = 0.90$. Fourth, the power of the RESET and HET does not improve as the sample size increases from $T = 50$ to $T = 200$, so these two tests applied to Eq. (11) may be inconsistent. Fifth, at every parameter combination, the power of LMF in detecting both autocorrelation and misspecification is now 100%.

Estimation by autocorrelation-robust methods Pagan and Hall (1983, pp. 206–209) suggest that the diagnostic tests can be made robust to model deficiencies other than those they were designed to detect, and their power can be improved, if the regression equation generating the diagnostics is estimated by robust methods rather than by OLS. We follow this suggestion and generate POY2, POY3, and POY4 from regressions estimated by the Cochrane–Orcutt (C–O) method. (The Hildreth–Lu method gives almost identical results.) Table 2 reports size and power of the RESET for $T = 50$. (Since the LMF and HET tests are generated by OLS regressions, their size and power are those reported in Table 1.)

First, notice that when $\rho = 0$, the actual size of RESET ranges from 5.92 to 10.28% and is significantly higher than the nominal size in every case. This is not surprising, since "correcting" for first-order autocorrelation where none exists causes the error term of the "corrected" equation to follow a first-order moving average scheme. But, as was discussed earlier, positive disturbance autocorrelation causes the RESET to over-reject, especially when the regressors are also positively autocorrelated. Compare this result with that from OLS at $\rho = 0$, where the size of RESET was approximately correct. Second, when $\rho = 0.5$ or $\rho = 0.95$, the estimated size of the RESET ranges from 3.98 to 13.26%, which is much better than that from OLS (0.26 to 68.90%). Third, the power of RESET is worse than in the OLS case.

[2] All the tables not reported here are available upon request.

Table 2 Rejection frequencies of the RESET generated from the C–O method (T=50)

			Correct model (unrestricted Eq. (11))			Misspecified null model (restrictions: $\beta_3=\beta_4=\beta_5=\beta_6=0$)		
φ	θ	ρ	POY2	POY3	POY4	POY2	POY3	POY4
0.00	0.00	0.00	6.40*	6.92*	7.44*	0.54	8.88	4.58
		0.50	5.08	5.80*	6.20*	0.02	5.40	2.20
		0.95	4.64	4.22*	4.48	0.00	2.84	0.68
	0.20	0.00	5.92*	6.96*	7.38*	1.66	1.22	0.48
		0.50	5.44	5.66*	5.96*	0.64	0.44	0.16
		0.95	4.30*	3.98*	4.28*	0.34	0.20	0.04
	0.90	0.00	5.96*	6.52*	7.50*	3.00	1.02	24.02
		0.50	5.10	5.76*	6.28*	1.86	0.26	21.62
		0.95	4.84	4.86	4.90	1.32	0.28	14.62
0.50	0.00	0.00	6.64*	7.04*	7.20*	0.18	0.02	0.04
		0.50	5.18	5.52	5.72*	0.00	0.00	0.00
		0.95	4.74	4.40	4.62	0.00	0.00	0.00
	0.20	0.00	6.60*	7.34*	8.06*	0.38	0.10	0.04
		0.50	5.60	5.80*	6.06*	0.06	0.02	0.00
		0.95	4.30*	4.54	4.74	0.04	0.00	0.00
	0.90	0.00	6.60*	7.54*	7.90*	4.86	1.36	0.64
		0.50	6.50*	7.00*	7.10*	2.76	0.32	0.04
		0.95	4.68	5.14	4.56	1.10	0.04	0.00
0.95	0.00	0.00	7.06*	7.54*	7.86*	0.00	0.24	0.06
		0.50	7.18*	6.64*	6.86*	0.00	0.04	0.00
		0.95	4.68	4.70	4.62	0.00	0.00	0.00
	0.20	0.00	7.34*	7.60*	8.38*	0.00	0.00	0.00
		0.50	7.10*	7.16*	6.84*	0.00	0.00	0.00
		0.95	5.26	5.08	4.52	0.00	0.00	0.00
	0.90	0.00	7.50*	9.38*	10.28*	0.10	0.00	0.00
		0.50	9.46*	12.52*	13.26*	0.02	0.00	0.00
		0.95	6.86*	7.00*	7.38*	0.00	0.00	0.00

(a) The rejection frequencies are given in percentages; (b) the entries under the heading "correct model" represent size, whereas those under "misspecified null model" represent power against the alternative of erroneously imposing the restrictions $\beta_3=\beta_4=\beta_5=\beta_6=0$; (c) the frequencies that fall outside the interval (4.40, 5.60) are regarded as significantly different from the nominal size (5%) and are marked by a star (*); (d) the rejection frequencies of LMF and HET are not reported in this table, because they are the same as those reported in Table 1.

Table 3 reports the results for $T = 200$. Now the size of the RESET ranges from 4 to 7% only, and does not depend on the values of φ, ρ, and θ. Thus, in large samples, using an autocorrelation-robust method to generate the RESET almost eliminates size distortion. Compare this result with that from OLS, where size distortion was greater when $T = 200$ than when $T = 50$. (The size and power of the LMF and HET tests reported in Table 3 are the same as those discussed earlier under OLS estimation for $T = 200$.)

Sensitivity analysis To see whether the above results can be generalized somewhat, we apply the five tests to Eq. (11) after erroneously imposing different sets of

Table 3 Rejection frequencies of the five diagnostic tests when T=200 and when the RESET is generated by the C–O method (the LMF and HET tests are generated by OLS)

			Correct model (unrestricted Eq. (11))					Misspecified null model (restrictions: $\beta_3=\beta_4=\beta_5=\beta_6=0$)				
φ	θ	ρ	POY2	POY3	POY4	LMF	HET	POY2	POY3	POY4	LMF	HET
0.00	0.00	0.00	5.06	5.76*	5.46	4.76	4.74	33.56	19.82	13.42	100.00	0.02
		0.50	4.78	5.32	5.44	100.00	4.12*	39.84	19.78	11.92	100.00	0.02
		0.95	4.20*	4.02*	4.24*	100.00	2.00*	46.50	20.32	9.64	100.00	0.52
	0.20	0.00	5.52	5.52	5.54	4.54	4.94	17.88	9.44	9.48	100.00	0.00
		0.50	4.72	4.88	4.58	100.00	4.70	14.74	5.98	5.56	100.00	0.12
		0.95	4.62	4.52	4.58	100.00	7.42*	13.78	3.18	2.98	100.00	5.68
	0.90	0.00	5.02	5.14	5.14	4.66	5.28	7.48	4.36	5.10	100.00	0.00
		0.50	5.26	5.34	5.42	100.00	6.78*	5.64	2.18	3.12	100.00	0.04
		0.95	4.86	4.82	5.26	100.00	17.04*	4.06	1.02	1.50	100.00	0.24
0.50	0.00	0.00	4.64	5.00	5.26	4.90	4.60	0.50	0.42	1.02	100.00	1.08
		0.50	5.00	4.90	5.08	100.00	5.92*	0.04	0.04	0.16	100.00	3.42
		0.95	5.26	4.62	4.50	100.00	9.42*	0.02	0.00	0.02	100.00	10.10
	0.20	0.00	5.22	5.34	5.42	5.52	5.02	5.18	1.90	8.76	100.00	0.04
		0.50	5.06	4.84	4.86	100.00	6.72*	2.92	0.54	4.32	100.00	0.18
		0.95	4.30*	4.52	5.08	100.00	11.54*	1.26	0.14	2.76	100.00	0.94
	0.90	0.00	5.64*	5.38	5.30	5.20	4.80	6.18	2.16	1.24	100.00	0.00
		0.50	5.52	5.50	5.06	100.00	8.72*	3.18	0.64	0.20	100.00	0.00
		0.95	4.48	4.60	4.88	100.00	26.10*	1.34	0.14	0.06	100.00	4.40
0.95	0.00	0.00	5.24	5.60	5.36	4.94	4.66	0.04	0.00	0.00	100.00	0.00
		0.50	6.02*	5.76*	5.62*	100.00	10.58*	0.00	0.00	0.00	100.00	0.00
		0.95	4.70	5.26	5.08	100.00	35.24*	0.00	0.00	0.00	100.00	2.02
	0.20	0.00	5.42	5.42	5.84*	5.26	5.42	0.18	0.14	0.02	100.00	0.00
		0.50	5.52	5.76*	5.94*	100.00	9.78*	0.00	0.00	0.00	100.00	0.00
		0.95	4.80	5.06	4.96	100.00	35.68*	0.00	0.00	0.00	100.00	1.72
	0.90	0.00	5.48	5.44	6.32*	4.66	4.52	0.34	0.00	0.00	100.00	0.00
		0.50	6.26*	6.96*	6.70*	100.00	11.20*	0.08	0.00	0.00	100.00	0.00
		0.95	5.34	5.38	5.70*	100.00	37.62*	0.00	0.00	0.00	100.00	2.32

(a) The rejection frequencies are given in percentages; (b) the frequencies that fall outside the interval (4.40, 5.60) are regarded as significantly different from the nominal size (5%) and are marked by a star (*); (c) the entries for LMF under the heading "correct model" represent size when ρ=0 and power against autocorrelation when ρ>0; (d) the entries under the heading "misspecified null model" represent power against the alternative of erroneously imposing the restrictions $\beta_3=\beta_4=\beta_5=\beta_6=0$.

restrictions.[3] Since the true model is still Eq. (11) with no restrictions imposed, the rejection frequencies reported in Table 1 under "correct model" remain valid, so we only consider the changes in the power of these tests against misspecification as we change the sets of the (wrong) restrictions imposed. Note at the outset that in every case the power of the RESET and HET does not differ noticeably from that reported

[3] In the first version of the paper, we only considered the omission of four lags of X_{1t}. We thank an anonymous referee of this journal for suggesting a generalization, so as to avoid conclusions that depend on a specific model.

in Table 1, so we will only refer to the changes in the power of the LMF test. Table 4 reports the results.

Briefly, when we erroneously impose the three restrictions $\beta_4 = \beta_5 = \beta_6 = 0$, the power of the LMF test against misspecification is virtually the same as that under the four restrictions $\beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ (see Table 1). The same is true when we impose only $\beta_5 = \beta_6 = 0$, but use $T = 200$. The power of LMF against misspecification falls markedly, however, when we impose $\beta_5 = \beta_6 = 0$ and use $T = 50$ as well as when we impose $\beta_6 = 0$. Note that imposing $\beta_6 = 0$ and using $T = 50$ reduces the power of LMF even against autocorrelation, especially when $\rho = 0.5$. (Compare the second to last column of Table 4 with that under "correct model" and LMF of Table 1.) This disappointing result is similar to that of Kiviet (1986, pp. 254–255) referred to earlier. But since the LMF test has good size properties, Kiviet (1986, p. 255) concludes that it "is a reliable model selection guideline" for dynamic models.

Table 4 Power of the LMF test under different sets of restrictions on the β's

Restrictions:			β_4=β_5=β_6=0		β_5=β_6=0		β_6=0	
φ	θ	ρ	T=50	T=200	T=50	T=200	T=50	T=200
0.00	0.00	0.00	99.82	100.00	55.30	100.00	0.82	0.70
		0.50	100.00	100.00	96.96	100.00	32.56	99.66
		0.95	100.00	100.00	99.98	100.00	98.16	100.00
	0.20	0.00	100.00	100.00	63.52	100.00	0.80	0.66
		0.50	100.00	100.00	97.94	100.00	42.74	99.78
		0.95	100.00	100.00	100.00	100.00	98.90	100.00
	0.90	0.00	99.68	100.00	75.78	100.00	0.66	0.84
		0.50	100.00	100.00	98.98	100.00	38.36	99.54
		0.95	100.00	100.00	100.00	100.00	98.56	100.00
0.50	0.00	0.00	99.84	100.00	15.14	100.00	1.26	1.02
		0.50	99.98	100.00	65.96	100.00	38.84	99.66
		0.95	100.00	100.00	98.26	100.00	98.24	100.00
	0.20	0.00	100.00	100.00	58.46	100.00	1.20	1.24
		0.50	100.00	100.00	93.76	100.00	30.38	99.80
		0.95	100.00	100.00	99.86	100.00	97.36	100.00
	0.90	0.00	100.00	100.00	47.12	100.00	0.96	0.80
		0.50	100.00	100.00	92.86	100.00	40.80	99.78
		0.95	100.00	100.00	99.84	100.00	98.60	100.00
0.95	0.00	0.00	100.00	100.00	47.72	100.00	0.38	1.00
		0.50	100.00	100.00	82.42	100.00	17.24	99.54
		0.95	100.00	100.00	97.52	100.00	91.72	100.00
	0.20	0.00	100.00	100.00	50.52	100.00	0.30	1.02
		0.50	100.00	100.00	85.50	100.00	20.62	99.28
		0.95	100.00	100.00	98.08	100.00	91.76	100.00
	0.90	0.00	100.00	100.00	51.28	100.00	0.44	0.58
		0.50	100.00	100.00	87.52	100.00	19.30	99.54
		0.95	100.00	100.00	97.62	100.00	91.54	100.00

5 Summary and conclusions

The paper examines the size and power of three widely used diagnostic tests applied to realistic distributed lag models: the RESET, a test for heteroscedasticity (HET), and a test for autocorrelation (LMF). If disturbance autocorrelation is absent ($\rho = 0$), these tests have the correct size approximately, but if $\rho > 0$, the RESET and HET tests are oversized, especially when the regressors are also positively autocorrelated. This finding agrees with those of Porter and Kashyap (1984) and Leung and Yu (2001). When a large sample is available, size distortion can be avoided almost entirely by using autocorrelation-robust methods.

With regards to power, the LMF test has high power (especially in large samples) to detect not only autocorrelation (provided that the model is correct), but also the erroneous omission of several lags of an explanatory variable, whereas the RESET and HET tests have no power to detect such misspecification errors. This result undermines the view that the RESET can be used "as a decent general check for model misspecification" (Leung and Yu 2001, p. 726).

References

Godfrey LG (1988) Misspecification tests in econometrics: the Lagrange multiplier principle and other approaches. Cambridge University Press, New York

Godfrey LG, McAleer M, McKenzie CR (1988) Variable addition and Lagrange multiplier tests for linear and logarithmic regression models. Rev Econ Stat 70:492–503

Godfrey LG, Orme CD (1994) The sensitivity of some general checks to omitted variables in the linear model. Int Econ Rev 35:489–506

Johnston J (1972) Econometric methods. 2nd Ed. McGraw-Hill, Tokyo

Kiviet JF (1986) On the rigour of some misspecification tests for modelling dynamic relationships. Rev of Econ Stud 53:241–261

Krämer W, Kiviet J, Breitung J (1990) The null distribution of the *F*-test in the linear regression model with autocorrelated disturbances. Statistica 50:503–509

Krämer W, Sonnberger H, Maurer J, Havlik P (1985) Diagnostic checking in practice. Rev Econ Stat 67:118–123

Leung SF, Yu S (2001) The sensitivity of the RESET tests to disturbance autocorrelation in regression residuals. Empir Econ 26:721–726

Pagan AR, Hall AD (1983) Diagnostic tests as residual analysis. Econom Rev 2:159–218

Porter RD, Kashyap AK (1984) Autocorrelation and the sensitivity of RESET. Econ Lett 14:229–233

Ramsey JB (1969) Tests for specification errors in classical linear least-squares regression analysis. J R Stat Soc, B 31:350–371

Thursby JG (1979) Alternative specification error tests: A comparative study. J Am Stat Assoc 74:222–225

Thursby JG (1989) A comparison of several specification error tests for a general alternative. Int Econ Rev 30:217–230

Thursby JG, Schmidt P (1977) Some properties of tests for specification error in a linear regression model. J Am Stat Assoc 72:635–641

Wooldridge JM (2003) Introductory econometrics: a modern approach. 2nd Ed., South-Western, Thomson