Behavior Research Methods
2007, 39 (3), 343-349

An improved portmanteau test for autocorrelated errors in interrupted time-series regression models

Bradley E. Huitema and Joseph W. McKean
Western Michigan University, Kalamazoo, Michigan

A new portmanteau test for autocorrelation among the errors of interrupted time-series regression models is proposed. Simulation results demonstrate that the inferential properties of the proposed Q_{H-M} test statistic are considerably more satisfactory than those of the well known Ljung–Box test and moderately better than those of the Box–Pierce test. These conclusions generally hold for a wide variety of autoregressive (AR), moving averages (MA), and ARMA error processes that are associated with time-series regression models of the form described in Huitema and McKean (2000a, 2000b).

Several approaches are available for the analysis of interrupted time-series (or time-series intervention) experimental designs. The simplest version of this design has two phases. The first phase (sometimes called the baseline or preintervention phase) has n_1 observations, and the second phase (postintervention) has n_2 observations. The purpose of the statistical analysis is to quantitatively describe and evaluate possible intervention effects.

Autoregressive integrated moving average (ARIMA) models and time-series regression models are frequently satisfactory for this purpose. In the case of ordinary least-squares (OLS) regression models, one assumes that the errors of the regression are independent. If this assumption is violated, the conventional hypothesis tests and confidence intervals on intervention effect coefficients are suspect (because they may be either conservative or liberal, depending on the sign of the autocorrelation), and an alternative method of regression analysis designed to correct for dependency of the errors should be considered. Common alternatives widely discussed in the econometric literature (see, e.g., Greene, 2000) and provided in popular software (such as SPSS 14.0) include several well-known versions of feasible generalized least-squares estimators. These include the Cochrane–Orcutt, Prais–Winston, and maximum-likelihood methods. The purpose of these methods is to transform the regression equation in order to remove first-order autocorrelation among the residuals of the fitted equation. Although some researchers routinely apply these methods without evaluating the need for them, there is a price to be paid for using such methods when the errors are not autocorrelated.

Simulation work has shown that the application of these methods in situations where there is no autocorrelation among the errors leads to an increase in Type I error for tests on intervention-effect coefficients (Huitema, McKean, & McKnight, 1994). Interestingly, the problem of distorted Type I error rates is also the main reason that these alternative estimation methods are recommended in the first place. Hence, there are two reasons why it is important to know whether the errors are autocorrelated. First, if the errors are not autocorrelated, OLS is the preferred method of estimation, and corrective methods should be avoided. Second, if the errors are autocorrelated, corrective methods are called for in order to maintain error rates at the nominal level. For these reasons, it is appropriate to carefully scrutinize the residuals of the fitted OLS regression equation for evidence of autocorrelated errors.

Many tests have been developed to identify autocorrelated errors; the classical Durbin–Watson (D–W) test (Durbin & Watson, 1950, 1951) is the best known. Two frequently cited problems with the D–W approach are that (1) the test statistic often falls in an inconclusive decision region, and (2) the test was not designed to identify autocorrelated errors that are generated by processes other than a first-order autoregressive model.

Discussions of the inconclusive region problem are contained in virtually all econometrics and regression textbooks (see, e.g., Greene, 2000; Johnston, 1984; Kutner, Nachtsheim, & Neter, 2004). The problem is that the exact critical value for the test is a function of the specific values in the design matrix. Because these values change with each application, the critical value is generally unknown, but it is bounded. Durbin and Watson provided tables (reproduced in many econometrics and regression texts) that contain these bounds. These tables contain both an upper and a lower critical value for each sample size and number of predictors. If the obtained D–W test statistic falls between the upper and lower critical values,

B. E. Huitema, brad.huitema@wmich.edu

the result is declared inconclusive. Solutions to the inconclusive region problem have been provided in the form of special purpose computer routines (e.g., White, 1993) that are not found in most software packages, as well as alternative tests (e.g., Huitema & McKean, 2000b).

The second perceived problem (i.e., inadequate sensitivity to errors generated by higher order error models) has led to recommendations (see, e.g., Shumway, 1988) to apply portmanteau autocorrelation tests that incorporate coefficients computed at many lags in the autocorrelation function. Two well-known methods that do this are the Box–Pierce (B–P) test (Box & Pierce, 1970) and the Ljung–Box (L–B) test (Ljung & Box, 1978); both tests were originally developed to evaluate errors of ARIMA models rather than errors of time-series regression models. Many popular software packages (such as Minitab, Version 14.0, and SPSS, Version 14.0) have implemented at least one of these tests. Although it has been frequently stated that the L–B test is superior to the B–P test in the context of conventional ARIMA models (see, e.g., Bowerman, O'Connell, & Koehler, 2005; Mills, 1990; Yaffee & McGee; 2000), we have recently shown (Huitema & McKean, 2007) that this is not true when these tests are applied to the residuals of certain regression models. Indeed, the L–B test has unacceptable Type I error properties regardless of sample size when it is used in the context of interrupted time-series regression models using design matrices of the form described in Huitema and McKean (2000a, 2000b).

The present article introduces and evaluates a portmanteau test for interrupted time-series regression models. It was developed to provide more satisfactory small sample properties than the conventional L–B and B–P portmanteau tests.

METHOD

Proposed Test Statistic $Q_{H–M}$

The expression for the proposed test statistic $Q_{H–M}$ is described below.

$$Q_{H–M} = \frac{N^3(N-1)}{(N-2)^2}\sum_{l=1}^{K}\frac{\left\{r_l + \left[\frac{P(N-l+1)}{N^2}\right]\right\}^2}{(N-l+1)},$$

where

K is the number of lags included in the test (generally between $N/15$ and $N/10$),
N is the total number of observations in the series,
P is the number of parameters in the regression model,
r_l is the autocorrelation coefficient computed on the residuals at lag-l; that is,

$$r_l = \frac{\sum_{t=1}^{N-l}(e_t)(e_{t+l})}{\sum_{t=1}^{N}e_t^2},$$

where e_t is the residual of the fitted model at time t, and $Q_{H–M}$ is distributed approximately as χ^2 with $df = K$.

Rationale for the $Q_{H–M}$ Test

The $Q_{H–M}$ statistic is a generalization of the the $z_{H–M}$ test statistic (Huitema & McKean, 2000b). The $z_{H–M}$ was designed to test H_0: $\phi_1 = 0$, whereas the $Q_{H–M}$ statistic was designed to test the joint hypothesis H_0: $\phi_1 = \phi_2 = \cdots = \phi_K = 0$. As is the case with the B–P and L–B tests, the test statistic $Q_{H–M}$ is essentially the sum of K ratios of squared autocorrelation estimates divided by their respective error variance estimates. The proposed test differs from the conventional portmanteau tests in both the method used to estimate the autocorrelation coefficients and the method used to estimate the error variances.

Previous work (Huitema & McKean, 2000b) has demonstrated that under the null hypothesis, the expected value of the conventional autocorrelation estimator r_l is negatively biased by a term that is a function of the number of parameters in the intervention regression model and the sample size. This term is added to each of the lag-1 through lag-K conventional autocorrelation estimates included in the test in order to reduce bias. It has also been demonstrated that the error variance estimators incorporated in the conventional B–P portmanteau test are positively biased with small samples (see Huitema & McKean, 1991). Slightly modified versions of the reduced bias variance estimator used in the $z_{H–M}$ test statistic (Huitema & McKean, 2000b) are incorporated in the proposed $Q_{H–M}$ test statistic. Because the modified autocorrelation and error variance estimators have been shown to be quite satisfactory in the $z_{H–M}$ test, we conjectured that the use of similar estimators in the joint-test context would lead to small sample performance that is superior to that of the B–P and L–B tests.

B–P and L–B Test Statistics

The expressions for the B–P and L–B test statistics are:

$$Q_{B–P} = N\sum_{l=1}^{K} r_l^2$$

and

$$Q_{L–B} = N(N+2)\sum_{l=1}^{K}(N-l)^{-1}r_l^2,$$

respectively. These test statistics are distributed approximately as chi square with $K - p - q$ degrees of freedom when they are applied to the residuals of ARMA models fitted to sample data.

One can see that $Q_{B–P}$ is simply N times the sum of the K squared autocorrelation coefficients, but that the expression for $Q_{L–B}$ appears to be considerably more complex. It turns out, however, that $Q_{L–B}$ is also N times the sum of K squared autocorrelation coefficients, but the autocorrelation coefficients are now redefined as

$$r_l^* = [(N+2)/(N-l)]^{1/2} r_l.$$

The logic for the modification is that the conventional formula for these coefficients contains only $N - l$ terms for the the autocovariance, whereas there are N terms for the variance, regardless of the lag. Hence, as the lag increases, the resulting coefficients are increasingly biased toward zero. The modification was designed to correct for this bias. Unfortunately, there are other sources of bias in both the conventional and modified coefficients when they are computed on the residuals of time-series regression models. The proposed test statistic was designed to reduce the large negative bias that is introduced by estimating multiple parameters in these models.

The proposed testing procedure involves both a new expression for the test statistic and degrees of freedom that differ from those used with the B–P and L–B methods. Previous work (see, e.g., Box & Pierce, 1970; Ljung, 1986; Ljung & Box, 1978) has shown that both $Q_{B–P}$ and $Q_{L–B}$ are asymptotically distributed approximately as χ_k^2 in the case where the ARMA parameters of the model are known. The proposed test statistic should also be distributed approximately as χ_k^2. Note in the expression for $Q_{H–M}$ that the ratio of the constants to the left of the summation will be 1 in the asymptotic case. Because K is finite and $\sqrt{N}r_l$ is bounded in probability under H_0, the difference between $Q_{L–B}$ and $Q_{H–M}$ goes to zero in probability under H_0. Hence, the asymptotic distributions of the three test statistics are the same.

In actual applications of the B–P and L–B tests with sample data, the parameters are estimated and the degrees of freedom are

$K - p - q$ (rather than K) for both tests. In the case of the H–M test, however, the degrees of freedom are K.

The reason Q_{B-P} and Q_{L-B} change *df* from K to $K - p - q$ when the ARMA parameters are estimated from the data is that bias is introduced to the estimates in the estimation process; this changes the distribution of the test statistic. Our approach is essentially to reduce the bias in the autocorrelation estimates rather than to modify the *df* to accommodate the bias in the coefficients; consequently, the distribution of the test statistic requires no *df* modification (i.e., $df = K$). The adequacy of this approach is demonstrated in the simulation results.

Intervention Model

The proposed test was developed to evaluate possible dependency among the errors of intervention regression models of the specific form recommended in Huitema and McKean (1998, 2000a, 2000b). These models provide estimates of level change and slope change for designs having two or more phases. The two-phase version of the model is written as follows:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 SC_t + \varepsilon_t,$$

where

Y_t is the dependent variable score at time t,
β_0 is the process intercept,
β_1, β_2, and β_3 are the process partial regression coefficients,
T_t is the value of the time variable T at time t,
D_t is the value of the level change dummy variable D (0 for the first phase and 1 for the second phase) at time t,
SC_t is the value of the slope change variable SC [defined as $T_t - (n_1 + 1)D_t$] at time t,
ε_t is the process error of the model at time t; the error process is assumed to be white noise.

The coefficients of direct experimental interest in this model are the level change coefficient, β_2, and the slope change coefficient, β_3. A thorough interpretation of these coefficients is described in Huitema and McKean (2000a). A description of the method used to specify the design matrices that are required to analyze time-series experiments with more than two phases can be found in Huitema and McKean (2000b).

Simulation Design

An extensive computer simulation experiment was carried out to evaluate the Type I error and power properties of the proposed test and competing procedures. The study included three levels of α (.01, .05, and .10), three test statistics (Q_{B-P}, Q_{H-M}, and Q_{L-B}), four lag depths (when sample size permitted), 30 error model combinations, and six sample sizes. Hence, results were obtained on approximately 6,480 (i.e., $3 \times 3 \times 4 \times 30 \times 6$) combinations. Descriptions of the specific error models included in the experiment and additional details regarding the simulation design follow.

The four-parameter intervention model described above was specified with the following parameter values: $\beta_0 = 0$, $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$. These parameters were then estimated for each simulated sample by applying OLS. The intervention point was specified to occur halfway through each simulated experiment, regardless of the total number of observations (N) in the simulated series.

The values of N (where $N = n_1 + n_2$) included in the experiment were $N = 20$, 30, 40, 50, 100, and 500. Most of the values of N were concentrated in the interval 20–50 because (1) many behavioral and social science experiments have only 10–25 observations per phase and (2) little is known about the small sample properties of portmanteau test statistics. The larger sizes were included to evaluate whether the performance of the test statistics conforms to theory.

Models for the Errors ε_t

The errors ε_t of the OLS regression model are generally assumed to be white noise, but alternative error models are likely to be more realistic in some studies. Details regarding the white noise model and the alternative models that were investigated follow.

1. White noise: $\varepsilon_t = [Y_t - (\beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 SC_t)]$, where the ε_ts are independent with equal variance at each time point and are normally distributed.
2. First-order autoregressive [AR(1)]: $\varepsilon_t = \phi_1 \varepsilon_{t-1} + u_t$, where ϕ_1 is the lag-1 autocorrelation of the process errors and ϕ_1 is the white noise disturbance at time t. The value of ϕ_1 was set at .2, .3, .4, .5, .7, and .9 for each sample size.
3. First-order moving average [MA(1)]: $\varepsilon_t = u_t + \theta_t u_{t-1}$, where θ_t was set at .2, .3, .4, .5, .7, and .9 for each sample size.
4. Sixth-order autoregressive [AR(6)]:

$$\varepsilon_t = \phi_j \sum_{j=1}^{6} \frac{7 - j}{6} \varepsilon_{t-j} + u_t,$$

where two patterns of process ϕ_js were specified as follows.

Pattern A: $\phi_1 = .150$, $\phi_2 = .125$, $\phi_3 = .100$, $\phi_4 = .075$, $\phi_5 = .050$, $\phi_6 = .025$

Pattern B: $\phi_1 = .200$, $\phi_2 = .167$, $\phi_3 = .133$, $\phi_4 = .100$, $\phi_5 = .067$, $\phi_6 = .033$

5. ARMA(1,1): $\varepsilon_t = \phi_1 \varepsilon_{t-1} + \theta_1 u_{t-1} + u_t$, where the parameters ($\phi_1$, θ_1) were set at the following 15 combinations.

(.20, −.50)
(.40, −.10)
(.40, −.70)
(.60, −.30)
(.60, −1.20)
(.80, −.60)
(.20, .10)
(−.20, −.10)
(−.20, .60)
(−.40, .10)
(−.40, .80)
(−.60, .40)
(−.60, 1.00)
(−.80, .60)
(−.80, 1.60)

The AR(1) and MA(1) error models were selected because they are relatively simple and appear to adequately represent much applied data. Because many procedures have been developed to accommodate AR(1) errors in regression models, the study of this structure is of considerable importance. High-order AR error models and mixed ARMA error models were included because the justification for portmanteau tests is to identify autocorrelation produced by these complex structures.

The values of the parameters included in the AR(1), MA(1), and AR(6) error models were chosen to avoid ceiling and floor effects (i.e., power values near 1 or 0) in simulations based on intermediate sample sizes. In the case of the ARMA(1,1) model, the parameter values were chosen to match those used by Andrews and Ploberger (1996). They chose values to correspond approximately to points on diagonal lines above and below the main diagonal with slope −1 in the (ϕ_1, θ_1) space. This method of selecting parameter values resulted in three combinations that contained one coefficient value falling outside the bounds of stationarity or invertibility. The results on these three combinations are not reported, but they were used in the process of checking the adequacy of the simulation methodology.

Specifically, the pattern of power results on the 15 ARMA combinations obtained by Andrews and Ploberger (1996) on the B–P test for the one sample size they studied ($N = 100$) was compared with

the pattern produced in the present study. Briefly, the average power in the two studies was .52 and .55 (at $K = 6$) for our results and for those of Andrews and Ploberger, respectively. The mean absolute deviation was .05. The somewhat lower values in our study are to be expected because our model had more parameters than did the model of Andrews and Ploberger; the direction of the difference is consistent with theory. The high agreement of the two patterns supports the adequacy of the simulation methodology.

Other approaches used to evaluate the simulation methodology include (1) comparisons of the parameter values specified in the simulations with the corresponding empirical mean values of the estimates, (2) comparisons of entire simulated error distributions against the corresponding theoretical distributions, (3) comparisons of empirical results with those of related studies on portmanteau tests (see, e.g., Newton, 1988, and Yaffee & McGee, 2000), and (4) comparisons of the results reported here (which were modeled using FORTRAN code) with those based on confirmatory results from S-Plus-based computer code. All criteria led to the conclusion that the following simulation approach was performing properly.

One thousand samples were generated for each combination of sample size, type of error model, and parameter value (or parameter combination values). Empirical Type I error rate and power functions were then computed. Each error series was started up with a standard normal variate. The first 300 observations of each generated series were ignored virtually to eliminate the correlation between the last observation of one simulated sample and the first observation of the next sample. The normal variates were obtained as discussed in Marsaglia and Bray (1964), using a FORTRAN generator written by Kahaner, Moler, and Nash (1989). The simulations were performed on a Sun SPARCstation I.

RESULTS

Empirical Type I Error

The empirical Type I error results for the proposed test are presented in Table 1 for the case where the nominal $\alpha = .05$ and $N = 20$, 30, 40, 50, 100, and 500. The Type I error rates for the B–P and L–B tests are also included in this table for comparative purposes. Results for other levels of α were also computed, but to conserve space, they are not shown.

One can see that the empirical error rates for the proposed test were closer to the nominal α value (.05) than were the rates for the B–P and L–B portmanteau tests. If we invoke the liberal version of the frequently used Bradley (1978) rule for acceptable test performance (viz., an empirical error rate that falls in the interval [.025–.075] for a 5% test), the rates for the proposed test were consistently acceptable for all sample sizes and lag depths (K) included in the study. The B–P test was generally acceptable for $K > 1$, but not for $K = 1$. In most cases, the L–B test was unacceptably liberal; for this reason, the power results for this test were not reported.

Table 1
Empirical Type I Error Associated With $Q_{H–M}$, $Q_{B–P}$, and $Q_{L–B}$ Test Statistics ($\alpha = .05$): White Noise Error Model

	$Q_{H–M}$		$Q_{B–P}$		$Q_{L–B}$	
N	$K = 1$	$K = *$	$K = 1$	$K = *$	$K = 1$	$K = *$
20	.03	.07	.11	.07	.15	.14
30	.04	.05	.09	.07	.11	.12
40	.05	.05	.09	.07	.10	.10
50	.05	.05	.08	.07	.09	.10
100	.05	.06	.06	.06	.07	.08
500	.05	.06	.05	.05	.05	.07

*$K = 3$ for $N = 20$, 30, 40, 50; $K = 8$ for $N = 100$; $K = 36$ for $N = 500$.

Table 2
Empirical Power Associated With $Q_{H–M}$ and $Q_{B–P}$ Test Statistics ($\alpha = .05$): AR(1) Error Model

	$Q_{H–M}$		$Q_{B–P}$	
Parameter Values	$K = 1$	$K = *$	$K = 1$	$K = *$
$N = 20$				
$\Phi_1 = .20$	.08	.06	.04	.06
$\Phi_1 = .30$	.13	.08	.03	.07
$\Phi_1 = .40$	.20	.11	.02	.09
$\Phi_1 = .50$	.26	.14	.03	.10
$\Phi_1 = .70$	.40	.23	.06	.14
$\Phi_1 = .90$	.49	.29	.09	.16
$N = 30$				
$\Phi_1 = .20$	.14	.08	.03	.07
$\Phi_1 = .30$	.25	.14	.06	.10
$\Phi_1 = .40$	.40	.25	.20	.13
$\Phi_1 = .50$	.53	.36	.22	.19
$\Phi_1 = .70$	.75	.60	.46	.35
$\Phi_1 = .90$	.86	.73	.62	.47
$N = 40$				
$\Phi_1 = .20$	.19	.12	.05	.09
$\Phi_1 = .30$	.36	.22	.13	.13
$\Phi_1 = .40$	.56	.39	.29	.22
$\Phi_1 = .50$	.74	.58	.47	.35
$\Phi_1 = .70$	.92	.83	.78	.65
$\Phi_1 = .90$	.97	.94	.91	.82
$N = 50$				
$\Phi_1 = .20$	.24	.16	.09	.10
$\Phi_1 = .30$	.47	.31	.24	.18
$\Phi_1 = .40$	.69	.52	.45	.33
$\Phi_1 = .50$	.86	.73	.68	.54
$\Phi_1 = .70$	.98	.94	.93	.86
$\Phi_1 = .90$	1.00	.99	.99	.96
$N = 100$				
$\Phi_1 = .20$	.48	.23	.31	.19
$\Phi_1 = .30$	.82	.51	.67	.44
$\Phi_1 = .40$	.96	.80	.91	.72
$\Phi_1 = .50$	1.00	.95	.99	.91
$\Phi_1 = .70$	1.00	1.00	1.00	1.00
$\Phi_1 = .90$	1.00	1.00	1.00	1.00
$N = 500$				
$\Phi_1 = .20$	.99	.67	.99	.62
$\Phi_1 = .30$	1.00	.98	1.00	.98
$\Phi_1 = .40$	1.00	1.00	1.00	1.00
$\Phi_1 = .50$	1.00	1.00	1.00	1.00
$\Phi_1 = .70$	1.00	1.00	1.00	1.00
$\Phi_1 = .90$	1.00	1.00	1.00	1.00

*$K = 3$ for $N = 20$, 30, 40, 50; $K = 8$ for $N = 100$; $K = 36$ for $N = 500$.

Empirical Power

AR(1) error model. Empirical power results are provided in Table 2 for the proposed portmanteau test statistic $Q_{H–M}$ and the conventional $Q_{B–P}$ test statistic for errors generated by an AR(1) error model. Notice that the proposed test was more powerful than the B–P test under almost all sample size and parameter value conditions listed in the table. Although the size of the $Q_{H–M}$ advantage was sometimes quite modest, there were other conditions under which the power of $Q_{H–M}$ was several times higher

than that of Q_{B-P}. The results for levels of α that are not shown here confirmed the pattern shown in Table 2. The overall power level increased as the nominal α level of the test increased, but the rank order of power of the tests remained the same, regardless of α.

MA(1) error model. Table 3 contains the results for the MA(1) error model. One can see that these results are very similar to those obtained for the AR(1) error model. For $N = 30$–500, the H–M portmanteau test provided higher power than did the B–P test, regardless of the level of the moving average coefficient that generated the errors. The only case in which the B–P portmanteau test appeared to be equal to the H–M test was with $N = 20$.

When low values (e.g., .20) of the moving average parameter were used to generate the errors, power was unacceptably low (<.80), even for $N = 500$ for both of the portmanteau ($K > 1$) tests. On the other hand, the relative performance of the different tests was generally consistent. There was no condition under which the $K = 1$ form (i.e., not portmanteau) of the H–M test was inferior to either form of the B–P test or the H–M test using $K > 1$. There were conditions under which the power of the $K = 1$ form of the H–M test was many times that of the corresponding form of the B–P test.

Table 3
Empirical Power Associated With Q_{H-M} and Q_{B-P} Test Statistics ($\alpha = .05$): MA(1) Error Model

	Q_{H-M}		Q_{B-P}	
Parameter Values	$K = 1$	$K = *$	$K = 1$	$K = *$
$N = 20$				
$\theta_1 = .20$	.08	.06	.04	.07
$\theta_1 = .30$	.14	.09	.02	.09
$\theta_1 = .40$	.21	.12	.01	.13
$\theta_1 = .50$	.28	.17	.02	.18
$\theta_1 = .70$	.42	.28	.04	.28
$\theta_1 = .90$	.49	.35	.06	.34
$N = 30$				
$\theta_1 = .20$	.13	.08	.03	.08
$\theta_1 = .30$	.25	.14	.05	.13
$\theta_1 = .40$	.38	.24	.09	.19
$\theta_1 = .50$	.51	.34	.15	.28
$\theta_1 = .70$	.69	.56	.30	.44
$\theta_1 = .90$	.77	.65	.37	.51
$N = 40$				
$\theta_1 = .20$	.18	.12	.05	.10
$\theta_1 = .30$	.34	.22	.11	.16
$\theta_1 = .40$	.52	.36	.23	.26
$\theta_1 = .50$	.69	.52	.35	.40
$\theta_1 = .70$	.85	.78	.58	.62
$\theta_1 = .90$	.90	.87	.66	.71
$N = 50$				
$\theta_1 = .20$	.23	.15	.09	.12
$\theta_1 = .30$	.45	.29	.20	.21
$\theta_1 = .40$	.65	.47	.36	.35
$\theta_1 = .50$	.80	.67	.55	.52
$\theta_1 = .70$	.93	.90	.78	.78
$\theta_1 = .90$	.96	.96	.85	.87
$N = 100$				
$\theta_1 = .20$	.46	.21	.29	.19
$\theta_1 = .30$	.79	.45	.62	.40
$\theta_1 = .40$	.94	.70	.86	.64
$\theta_1 = .50$	.99	.89	.96	.83
$\theta_1 = .70$	1.00	.99	1.00	.98
$\theta_1 = .90$	1.00	1.00	1.00	.99
$N = 500$				
$\theta_1 = .20$	.99	.62	.99	.58
$\theta_1 = .30$	1.00	.97	1.00	.95
$\theta_1 = .40$	1.00	1.00	1.00	1.00
$\theta_1 = .50$	1.00	1.00	1.00	1.00
$\theta_1 = .70$	1.00	1.00	1.00	1.00
$\theta_1 = .90$	1.00	1.00	1.00	1.00

*$K = 3$ for $N = 20, 30, 40, 50$; $K = 8$ for $N = 100$; $K = 36$ for $N = 500$.

Table 4
Empirical Power Associated With Q_{H-M} and Q_{B-P} Test Statistics ($\alpha = .05$): AR(6) Error Model

	Q_{H-M}		Q_{B-P}	
N	$K = 1$	$K = *$	$K = 1$	$K = *$
Pattern A. Parameter Values: $\Phi_1 = .150$, $\Phi_2 = .125$, $\Phi_3 = .100$, $\Phi_4 = .075$, $\Phi_5 = .050$, $\Phi_6 = .025$.				
20	.04	.07	.11	.07
30	.06	.06	.07	.06
40	.08	.08	.06	.05
50	.12	.12	.07	.06
100	.32	.33	.21	.19
500	.98	.98	.98	.98
Pattern B. Parameter Values: $\Phi_1 = .200$, $\Phi_2 = .167$, $\Phi_3 = .133$, $\Phi_4 = .100$, $\Phi_5 = .067$, $\Phi_6 = .033$.				
20	.04	.07	.10	.07
30	.06	.06	.07	.06
40	.10	.09	.06	.05
50	.15	.16	.08	.07
100	.52	.57	.40	.39
500	1.00	1.00	1.00	1.00

*$K = 3$ for $N = 20, 30, 40, 50$; $K = 8$ for $N = 100$; $K = 36$ for $N = 500$.

AR(6) error model. Table 4 displays the results under the sixth-order autoregressive error model. One can see that results are presented separately for each of the two patterns (A and B) of autoregressive parameters. Notice that (1) both forms of the H–M test generally produced higher power than did the corresponding B–P test, and (2) there was very little difference between the performance of the $K = 1$ and $K > 1$ forms of these tests.

ARMA(1,1) error model. Power results for the data generated under the first-order autoregressive-moving average error model are presented in Table 5. Each row in the table is associated with a unique combination of autoregressive (ϕ_1) and moving average (θ_1) parameter values. One can see that the overall performance across the whole collection of parameter combinations is about the same for the H–M and B–P methods. Interestingly, this is true whether lag depth of the test is set at 1 or a higher value.

DISCUSSION

The proposed test was designed to have better inferential properties than the traditional B–P and L–B portmanteau approaches for testing the independence assumption. This goal was met; the H–M portmanteau test is generally the method of choice among the portmanteau tests evaluated. However, the conjecture that portmanteau tests (based on autocorrelation coefficients computed at many lags) are

Table 5
Empirical Power Associated With Q_{H-M} and Q_{B-P} Test Statistics ($\alpha = .05$): ARMA(1,1) Error Model

	Q_{H-M}		Q_{B-P}	
Φ_1 and θ_1 Parameter Values	$K = 1$	$K = *$	$K = 1$	$K = *$
$N = 20$				
(.20, −.50)	.05	.14	.03	.14
(.40, −.10)	.12	.14	.30	.14
(.40, −.70)	.04	.19	.22	.21
(.60, −.30)	.10	.07	.04	.06
(.80, −.60)	.05	.06	.08	.06
(.20, .10)	.13	.15	.02	.16
(−.20, −.10)	.09	.21	.39	.20
(−.20, .60)	.19	.12	.13	.14
(−.40, .10)	.14	.25	.45	.25
(−.40, .80)	.16	.12	.01	.14
(−.60, .40)	.10	.20	.36	.22
(−.80, .60)	.15	.29	.42	.31
$N = 50$				
(.20, −.50)	.29	.31	.52	.52
(.40, −.10)	.46	.32	.24	.18
(.40, −.70)	.20	.15	.41	.29
(.60, −.30)	.44	.33	.24	.16
(.80, −.60)	.20	.17	.08	.07
(.20, .10)	.46	.30	.23	.19
(−.20, −.10)	.43	.33	.66	.44
(−.20, .60)	.59	.45	.30	.34
(−.40, .10)	.52	.54	.72	.66
(−.40, .80)	.47	.59	.20	.48
(−.60, .40)	.32	.32	.51	.38
(−.80, .60)	.40	.48	.59	.55
$N = 100$				
(.20, −.50)	.66	.37	.80	.45
(.40, −.10)	.81	.53	.68	.44
(.40, −.70)	.53	.30	.69	.40
(.60, −.30)	.82	.61	.71	.50
(.80, −.60)	.48	.37	.34	.24
(.20, .10)	.80	.48	.67	.41
(−.20, −.10)	.80	.52	.89	.58
(−.20, .60)	.90	.65	.79	.60
(−.40, .10)	.85	.62	.92	.66
(−.40, .80)	.80	.59	.63	.55
(−.60, .40)	.58	.46	.71	.49
(−.80, .60)	.68	.67	.78	.69
$N = 500$				
(.20, −.50)	1.00	.95	1.00	.96
(.40, −.10)	1.00	.99	1.00	.99
(.40, −.70)	1.00	.95	1.00	.96
(.60, −.30)	1.00	1.00	1.00	1.00
(.80, −.60)	1.00	.98	1.00	.97
(.20, .10)	1.00	.98	1.00	.97
(−.20, −.10)	1.00	.98	1.00	.98
(−.20, .60)	1.00	1.00	1.00	1.00
(−.40, .10)	1.00	.97	1.00	.97
(−.40, .80)	1.00	1.00	1.00	1.00
(−.60, .40)	1.00	.91	1.00	.91
(−.80, .60)	1.00	.99	1.00	.99

*$K = 3$ for $N = 20, 50$; $K = 8$ for $N = 100$; $K = 36$ for $N = 500$.

more powerful than tests based on only the lag-1 coefficient is not supported. For the typical application in which the researcher simply wants to know if there is convincing evidence of dependency among the errors, the most reasonable approach is to use a test that focuses on the lag-1 coefficient. Either the classic D–W test or the Q_{H-M} test using $K = 1$ (which is equivalent to the z_{H-M} test) is a better choice in this situation. The power of these tests is higher than the power of any of the portmanteau ($K > 1$) tests under a wide variety of commonly encountered error models. The power advantage of the simpler tests is generally substantial.

This result may be surprising to those who believe that the additional information incorporated into portmanteau tests should increase power. One reason this was not found can be traced to the specific hypotheses associated with the different tests. As is the case in many areas of statistics, a penalty is paid for joint tests. The D–W and Q_{H-M} (using $K = 1$) approaches focus on only one parameter in testing H_0: $\phi_1 = 0$, whereas the portmanteau methods test the joint hypothesis H_0: $\phi_1 = \phi_2 = \cdots = \phi_K = 0$ (where $K > 1$). Because the error models most frequently encountered in practice (other than white noise) have a larger autocorrelation coefficient at lag-1 than at other lags, it is reasonable for one to expect the more focused tests to have higher power than any of the portmanteau tests when sample sizes are not large.

Although the results in this article are based on a four-parameter regression model, it is likely that the advantage of the proposed portmanteau test over the B–P test would be amplified in the case of regression models having more parameters. If, for example, we have a design with four phases rather than two, the number of parameters in the regression model will be eight rather than four. As the number of parameters increases, the amount of negative bias in the autocorrelation estimates included in conventional portmanteau tests will also increase. This bias will increase the Type I error rate of the test.

Evidence of this effect was observed in the present study. The empirical error rate for the B–P test was actually higher than power (against low values of AR or MA parameters) in the case of very small samples. This result occurred because negative bias in the autocorrelation coefficients produced squared autocorrelation estimates that were higher under the null hypothesis than the corresponding coefficients that were computed when there was positive dependency among the errors. In the case of larger samples, however, the power advantage of the simpler tests increases as a function of the ratio of the number of parameters in the regression model to N.

Because no test for autocorrelated errors has adequate power with small samples (say, $N \leq 50$), when α is set at conventional levels, the researcher needs a strategy for this situation. We recommend that the following alternatives be considered: (1) Set α for the autocorrelation test at .10 (rather than at the conventional .05) and proceed as usual; (2) instead of testing for autocorrelated errors, use a model selection method such as the Schwarz Bayesian criterion (Schwarz, 1978) to decide whether error models with ARMA parameters are better than the white noise model; and (3) ignore both formal testing and model selection criteria and instead use a method for estimating the intervention model that has better small sample performance than any of the conventional feasible generalized least-squares estimators (see, e.g., McKnight, McKean, & Huitema, 2000). If the third approach is followed, care should be taken to avoid methods that have been shown to be invalid (see Huitema, 2004; Huitema, McKean, & Laraway, in press).

AUTHOR NOTE

Correspondence concerning this article should be addressed to B. E. Huitema, Department of Psychology, Western Michigan University, Kalamazoo, MI 49008 (e-mail: brad.huitema@wmich.edu).

REFERENCES

ANDREWS, D. W. K., & PLOBERGER, W. (1996). Testing for serial correlation against an ARMA(1,1) process. *Journal of the American Statistical Association*, **91**, 1331-1342.

BOWERMAN, B. L., O'CONNELL, R. T., & KOEHLER, A. B. (2005). *Forecasting, time series, and regression: An applied approach* (4th ed.). Belmont, CA: Thomson, Brooks/Cole.

BOX, G. E. P., & PIERCE, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, **65**, 1509-1526.

BRADLEY, J. V. (1978). Robustness? *British Journal of Mathematical & Statistical Psychology*, **31**, 144-152.

DURBIN, J., & WATSON, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, **37**, 409-428.

DURBIN, J., & WATSON, G. S. (1951). Testing for serial correlation in least squares regression: II. *Biometrika*, **38**, 159-178.

GREENE, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

HUITEMA, B. E. (2004). Analysis of interrupted time-series experiments using ITSE: A critique. *Understanding Statistics: Statistical Issues in Psychology, Education, & the Social Sciences*, **3**, 27-46.

HUITEMA, B. E., & MCKEAN, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin*, **110**, 291-304.

HUITEMA, B. E., & MCKEAN, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, **3**, 104-116.

HUITEMA, B. E., & MCKEAN, J. W. (2000a). Design specification issues in time-series intervention models. *Educational & Psychological Measurement*, **60**, 38-58.

HUITEMA, B. E., & MCKEAN, J. W. (2000b). A simple and powerful test for autocorrelated errors in OLS intervention models. *Psychological Reports*, **87**, 3-20.

HUITEMA, B. E., & MCKEAN, J. W. (2007). Identifying autocorrelation generated by various error processes in interrupted time-series regression designs: A comparison of AR1 and portmanteau tests. *Educational & Psychological Measurement*, **67**, 447-459.

HUITEMA, B. E., MCKEAN, J. W., & LARAWAY, S. (in press). Time-series intervention analysis using ITSACORR: Fatal flaws. *Journal of Modern Applied Statistical Methods*.

HUITEMA, B. E., MCKEAN, J. W., & MCKNIGHT, S. D. (1994, August). *Small-sample time-series intervention analysis: Problems and solutions*. Paper presented at the meeting of the American Psychological Association, Los Angeles.

JOHNSTON, J. (1984). *Econometric methods* (3rd ed.). New York: McGraw-Hill.

KAHANER, D., MOLER, C., & NASH, S. (1989). *Numerical methods and software*. Englewood Cliffs, NJ: Prentice Hall.

KUTNER, M. H., NACHTSHEIM, C. J., & NETER, J. (2004). *Applied linear regression models* (4th ed.). New York: McGraw-Hill Irwin.

LJUNG, G. M. (1986). Diagnostic testing of univariate time series models. *Biometrika*, **73**, 725-730.

LJUNG, G. M., & BOX, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, **65**, 297-303.

MARSAGLIA, G., & BRAY, T. A. (1964). A convenient method for generating normal variables. *SIAM Review*, **6**, 260-264.

MCKNIGHT, S., MCKEAN, J. W., & HUITEMA, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods*, **5**, 87-101.

MILLS, T. C. (1990). *Time series techniques for economists*. Cambridge: Cambridge University Press.

NEWTON, H. J. (1988). *TIMESLAB: A Time Series Analysis Laboratory*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

SHUMWAY, R. H. (1988). *Applied statistical time series analysis*. Englewood Cliffs, NJ: Prentice Hall.

WHITE, K. (1993). *SHAZAM* (Version 7) [Computer software]. Vancouver: University of British Columbia, Department of Economics.

YAFFEE, R. A., & MCGEE, M. (2000). *Introduction to time series analysis and forecasting: With applications of SAS and SPSS*. San Diego: Academic Press.

(Manuscript received June 29, 2005;
revision accepted for publication February 27, 2006.)