

A New Procedure for Multiple Testing of Econometric Models

Maxwell L. King¹, Xibin Zhang, Muhammad Akram

Department of Econometrics and Business Statistics, Monash University, Australia

3 February 2010

Abstract: A significant role for hypothesis testing in econometrics involves diagnostic checking. When checking the adequacy of a chosen model, researchers almost always employ a range of diagnostic tests, each of which is designed to detect a particular form of model inadequacy. A major problem is how best to control the overall probability of rejecting the model when it is true and multiple test statistics are used. This paper presents a new multiple testing procedure, which involves checking whether the calculated values of the diagnostic statistics are consistent with the postulated model being true. This is done through a combination of bootstrapping to obtain a multivariate kernel density estimator of the joint density of the test statistics under the null hypothesis and Monte Carlo simulations to obtain a p value using this kernel density. The proposed testing procedure is applied to tests for autocorrelation in an observed time series, for normality, and for model misspecification. We find that our testing procedure has correct or nearly correct sizes and good powers, particular for more complicated testing problems. We believe it is the first good method for calculating the overall p value for a vector of test statistics based on simulation.

Key words: bootstrapping; multivariate kernel density; normality; serial correlation; test vector

JEL classification: C01, C12, C14.

¹Corresponding Author. Email: max.king@adm.monash.edu.au; telephone: +61-3-99052449; address: Monash University, Victoria 3800, Australia.

1 Introduction

Statistical hypothesis testing is an extremely important technique in the practice of econometrics, particularly with respect to diagnostic checking of model specification. This is how econometricians are best able to combat the severe problem of uncertainty in model specification. Such testing procedures need to be as accurate as possible due to constraints on data availability. Fortunately, advances in computer power and simulation based methods have allowed greater scope in the design of high quality tests. The purpose of any test is to accurately control the probability of wrongly rejecting the null hypothesis (known as the size of the test), while at the same time ensuring a high probability of correctly rejecting the null hypothesis (known as the power of the test).

There is a very large literature on diagnostic testing of all kinds of econometric models. Therefore, in order to check the adequacy of a chosen model, researchers can apply a range of diagnostic tests, each of which is designed to detect a particular form of model inadequacy. A major problem is how best to control the overall probability of rejecting the model when it is true. For example, five statistically independent tests applied at the 5% level will result in a 22.6% chance of at least one rejection when the null hypothesis model is true. Of course, it is unlikely that five diagnostic tests applied to the same model will be mutually independent, so in actual fact, this probability could be higher or lower than 22.6%. The major issue is how we should conduct these tests in order to control the overall probability of rejecting the model when it is true.

The aim of this paper is to develop a new procedure for testing based on multiple test statistics in a way that controls the overall probability of a false rejection. Let \mathbf{y} denote a vector of T observations. A typical approach to hypothesis testing is to construct the critical region via a test statistic denoted by $t(\mathbf{y}) : R^T \rightarrow R$, which is a mapping from the T -dimensional sample space to the real line and follows, or at least asymptotically follows, a known distribution under the null hypothesis. If the sample falls in the critical region, the null hypothesis is rejected. Multiple hypothesis testing is the testing of two or more separate parameters or hypotheses simultaneously.

Often, each parameter or hypothesis being tested gives rise to its own statistic which is then combined with the other statistics to form one test statistic in a way that gives a convenient asymptotic distribution under the null hypothesis. Good examples are multidimensional Wald and Lagrange multiplier (LM) tests. The method of combination of the component statistics can be arbitrary and may involve an estimate of the asymptotic covariance matrix of the component statistics that can be a rather poor estimate of the actual covariance matrix. These problems can affect the small-sample size and power properties of the resultant test.

This paper proposes an alternative approach to multiple hypothesis testing based on a vector of test statistics, $t(\mathbf{y}): R^T \rightarrow R^d$ ($T > d$). It assumes that each of the elements of $t(\mathbf{y})$ has been chosen because individually they have good power to detect a particular deviation from the null hypothesis. It is further assumed that collectively $t(\mathbf{y})$ provides a good summary of the evidence contained in \mathbf{y} that might point to the null hypothesis being false. The approach involves asking the question based on the observed value of $t(\mathbf{y})$, do we think the null hypothesis is true? If we know the joint density function for $t(\mathbf{y})$ under the null hypothesis, then following Hyndman (1996) we can calculate the p value for the observed value of $t(\mathbf{y})$. Typically we do not know this joint density function. Our approach is to simulate independent values of $t(\mathbf{y})$ under the null hypothesis and then use a multivariate kernel density estimator to estimate the density. The contribution we make in this paper can be viewed as a way to use simulation methods to (approximately) control the probability of falsely rejecting the null hypothesis based on a vector of test statistics. How this can be done for a single statistic is well understood. To the best of our knowledge, our test procedure is the first general method for calculating p -values based on simulation.

The rest of the paper is organized as follows. Section 2 presents the new testing procedure for invariant test statistics. In Section 3, we examine the performance of the new testing procedure through Monte Carlo simulations, where we provide comparisons of the performances in terms of size and power between our testing procedure and some other commonly used test statistics. In Section 4, we present the testing procedure for non-invariant test statistics, where a bootstrapping procedure is used. Section 5 briefly describes the information matrix test and its limitations. We present a Monte Carlo simulation study of the new testing procedures applied to the information

matrix test in Section 6. In Section 7, we analyze the simulation results. Section 8 concludes the paper.

2 The Testing Procedure for Invariant Test Statistics

2.1 Testing Procedure

We shall begin by first describing the main ideas behind our new testing procedure. Assume that we are interested in testing the null hypothesis that the $T \times 1$ vector of observations \mathbf{y} has a particular data generating process using d test statistics denoted as t_i , for $i = 1, 2, \dots, d$. Let $\mathbf{t} = (t_1, t_2, \dots, t_d)'$ represent the $d \times 1$ vector of the test statistics called the test vector hereafter. At the moment, we assume that each of the component tests is a two-sided test based on accepting the null hypothesis if

$$c_{1i} < t_i < c_{2i},$$

where c_{1i} and c_{2i} are critical values, for $i = 1, 2, \dots, d$. We also assume t_i , for $i = 1, 2, \dots, d$, are similar tests in the sense that their distribution under the null hypothesis is invariant to nuisance parameters. Let $\hat{\mathbf{t}}$ denote the calculated value of the test vector \mathbf{t} using the available data.

Essentially, we wish to ask the question, is the calculated value of our test vector consistent with the null hypothesis being true? The p value is a useful tool for answering this question. It is defined as the probability under the null hypothesis of finding a value of the test vector as extreme as or more extreme than the value we have found from the data, namely $\hat{\mathbf{t}}$. Thus, if we have the joint density of \mathbf{t} denoted by $f(\mathbf{t})$, under the null hypothesis, the p value of the test vector is the probability of obtaining a value of \mathbf{t} such that $f(\mathbf{t}) < f(\hat{\mathbf{t}})$ holds. Once calculated, the p value can be used to conduct the test at any level of significance. For example, at the 5% significance level, if the p value is less than 0.05 then the null hypothesis is rejected. Otherwise, it cannot be rejected. The resultant acceptance region is optimal in the sense that by its construction, it is the smallest 95% acceptance region in the d -dimensional sample space of \mathbf{t} . If we believe $t(\mathbf{y})$ provides a good summary of the evidence contained in \mathbf{y} that might point to the null hypothesis

being false then this is a desirable property to have.

Typically the d -dimensional density $f(\mathbf{t})$ is unknown. We can estimate it by applying a multivariate kernel density estimator to a sample of independent drawings from $f(\mathbf{t})$ which can be obtained by repeatedly simulating the data generating process under the null hypothesis and then calculating \mathbf{t} for each simulated data set. Let $\{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_m\}$ denote such a sample. The general form of the kernel density estimator of \mathbf{t} is given by

$$\hat{f}_H(\mathbf{t}) = \frac{1}{m} \sum_{i=1}^m |H|^{-1/2} K(H^{-1/2}(\mathbf{t} - \mathbf{t}_i)), \quad (1)$$

where $K(\cdot)$ is a kernel function, and H is a positive definite matrix of bandwidths known as the bandwidth matrix (see Scott, 1992; Wand and Jones, 1995; among others).

There are two ways in which the new testing procedure can be implemented in practice. The first involves two separate rounds of simulation as follows:

- (i) Based on the data under test, calculate $\hat{\mathbf{t}}$.
- (ii) Using any convenient form of the data generating process under the null hypothesis, simulate the model m times and calculate m independent values of \mathbf{t} denoted as $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_m$.
- (iii) Use the sample generated in (ii) to estimate the joint density $f(\mathbf{t})$ by $\hat{f}_H(\mathbf{t})$ via (1).
- (iv) Repeat (ii) to generate a second sample of n values of \mathbf{t} denoted as $\mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \dots, \mathbf{t}^{(n)}$, which are independent of those originally calculated in step (ii).
- (v) Use this second sample to calculate $\hat{f}_H(\mathbf{t}^{(i)})$, for $i = 1, 2, \dots, n$. The p value of the joint test is estimated by the relative frequency for which $\hat{f}_H(\mathbf{t}^{(i)}) < \hat{f}_H(\hat{\mathbf{t}})$ holds.

The second test procedure involves only one round of simulation. After completing steps (i) and (ii) above, the remaining steps are as follows:

(iii') Use the sample generated in (ii) to estimate the joint density $f(\mathbf{t})$ at $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_m$ by the leave-one-out kernel density estimator

$$\hat{g}_H(\mathbf{t}_i) = \frac{1}{m-1} \sum_{\substack{j=1 \\ j \neq i}}^m |H|^{-1/2} K(H^{-1/2}(\mathbf{t}_i - \mathbf{t}_j)). \quad (2)$$

(iv') The p value of the joint test is estimated by the relative frequency for which $\hat{g}_H(\mathbf{t}_i) < \hat{f}_H(\hat{\mathbf{t}})$ holds.

2.2 Bandwidth Selection

The multivariate kernel density estimator depends on the choice of a bandwidth matrix and the choice of a kernel function. It is generally accepted in the statistical literature that the performance of the kernel density estimator is mainly determined by the bandwidth matrix, and only in a minor way by the choice of a kernel function. The bandwidth matrix can be either a full matrix or a diagonal matrix. A full bandwidth matrix is able to incorporate any possible correlation between any pair of the d dimensions. However, the number of nonzero bandwidths to be estimated in a full bandwidth matrix grows dramatically as d increases. Consequently, a full bandwidth matrix encounters more computing complexity in selecting a bandwidth matrix that is optimal with respect to a chosen criterion than a diagonal bandwidth matrix does. As discussed by Wand and Jones (1993) in the situation of the bivariate kernel density estimation, a diagonal bandwidth matrix allows for the flexibility of choosing a different bandwidth in each dimension and is often appropriate. Therefore, we use a diagonal bandwidth matrix in this new testing procedure, where the bandwidth matrix is denoted as $H = \text{diag}\{h_1^2, h_2^2, \dots, h_d^2\}$.

According to Scott (1992) and Bowman and Azzalini (1997), when data are observed from the multivariate normal density and the diagonal bandwidth matrix is used, the optimal bandwidth matrix that minimizes the MISE between the true density and its estimated density can be approximated by

$$h_i = \sigma_i \left\{ \frac{4}{(d+2)n} \right\}^{1/(d+4)},$$

for $i = 1, 2, \dots, d$, where σ_i is the standard deviation of the i th variate and can be replaced by its sample estimator in practice. We call this the normal reference rule (NRR) which is also known as the rule-of-thumb method in the literature. This bandwidth selection method is often used in many applications of multivariate kernel density estimation in the absence of any other practical bandwidth selection methods, despite the fact that the data might not be Gaussian.

Zhang, King and Hyndman (2006) presented a Bayesian sampling algorithm to estimate the bandwidth matrix in multivariate kernel density estimation. The bandwidth matrix chosen through this sampling algorithm tends to produce a more accurate density estimator than that chosen through the NRR. However, the Bayesian bandwidth selector is far more computational costly than the NRR. A general guideline for selecting one of the two bandwidth selectors is as follows. When the required computing time is not of serious concern in the testing procedure, one may use the Bayesian bandwidth selector. Otherwise, one may use the NRR to choose bandwidths.

As the performance of the kernel density estimator is only slightly affected by the choice of a kernel function, we will not investigate the issue about the choice of kernel. Throughout this paper, we use the product of d univariate Gaussian kernels in the kernel density estimator of $f(\mathbf{t})$.

3 Monte Carlo Experiments for Invariant Test Statistics

We conducted two separate Monte Carlo experiments in order to study the small sample size and power performance of the new test procedure. The testing problems involved are (i) testing for autocorrelation in a stationary time series; and (ii) testing for normality in a simple random sample. In each case we compared the performance of two versions of our test procedure, one using the Bayesian bandwidth selector and the other using the NRR bandwidth vector with an established benchmark test.

As simulations of simulations can be very time consuming, we used the following approach to estimate the size and power of the test procedure.

- (a) Simulate a convenient version of the data generating process under the null hypothesis and calculate a simple random sample of m values of \mathbf{t} , denoted as $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_m$.
- (b) For the particular choice of bandwidth matrix, compute the value of the leave-one-out kernel density $\hat{g}_H(\mathbf{t}_i)$, for $i = 1, 2, \dots, m$.
- (c) Order the values of $\hat{g}_H(\mathbf{t}_i)$ from the lowest to highest and for a test at the α percent significance level, find the α percentile of these $\hat{g}_H(\mathbf{t}_i)$ values. Denote the α percentile value as $\hat{g}_\alpha(\mathbf{t})$. Essentially $\hat{g}_\alpha(\mathbf{t})$ acts as a critical value for a test with $\hat{f}_H(\hat{\mathbf{t}})$ as the test statistic.
- (d) Simulate the data generating process under which size or power is to be estimated and calculate a second simple random sample of n values of \mathbf{t} denoted as $\mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \dots, \mathbf{t}^{(n)}$. The estimated probability of rejection of the null hypothesis is the relative frequency that

$$\hat{f}_H(\mathbf{t}^{(i)}) < \hat{g}_\alpha(\mathbf{t}),$$

holds for $i = 1, 2, \dots, n$.

Step (d) is repeated for a range of data generating processes using the same set of model disturbances to allow good comparability.

3.1 Testing for Serial Correlation of Unknown Order and Form

3.1.1 The Experimental Design

The first Monte Carlo experiment involves the classical problem of testing the null hypothesis that an observed time series is white noise against the alternative that it contains serial correlation of unknown order and form (see King, 1987). The null hypothesis is of the form

$$y_t = \mu + \varepsilon_t, \quad \text{for } t = 1, 2, \dots, T, \tag{3}$$

where μ is an unknown parameter and ε_t are independent and identically distributed as $N(0, \sigma^2)$. The alternative is that there is some serial correlation in ε_t and it is assumed that it can best be

detected by examining r_k , the k th order autocorrelation coefficient, for $k = 1, 2, \dots, d$, where

$$r_k = \frac{\sum_{t=k+1}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{\sum_{t=1}^T \hat{\varepsilon}_t^2},$$

in which $\hat{\varepsilon}_t$, for $t = 1, 2, \dots, T$, are the ordinary least squares (OLS) residuals from fitting (3) to the observed time series. In other words,

$$t_i = r_i, \quad \text{for } i = 1, 2, \dots, d,$$

in this problem. Note that r_i is invariant to the values of μ and σ^2 under (3).

A standard testing procedure for this problem is to use the Portmanteau test proposed by Box and Pierce (1970) and extended by Ljung and Box (1978). It involves rejecting the null hypothesis for large values of the Portmanteau test statistic given by

$$Q_d = T(T+2) \sum_{k=1}^d \frac{r_k^2}{T-k}. \quad (4)$$

The Monte Carlo experiment involved comparing sizes and powers of the Portmanteau test based on (4) applied using simulated critical values with two forms of the new procedure, the first using NRR bandwidth parameters and the second using MCMC bandwidth parameters for $d = 4$ and $d = 6$, respectively.

Sizes were calculated by simulating the data generating process using (3) with $\varepsilon_t \sim IN(0, 1)$ and $\mu = 1$. Powers were calculated for four different data generating processes for ε_t in (3), these being ε_t generated by

- (i) the stationary first-order autocorrelation process (AR(1)) given by

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t,$$

or equivalently $(1 - \rho L)\varepsilon_t = u_t$, where $\rho = 0.25$, $u_t \sim IN(0, 1)$ and L is the lag operator;

- (ii) the stationary second-order autocorrelation process (AR(2)) given by

$$(1 - \rho_1 L)(1 - \rho_2 L)\varepsilon_t = u_t$$

with $(\rho_1, \rho_2) = (0.05, 0.1)$ and $(0.05, 0.2)$ and $u_t \sim IN(0, 1)$;

(iii) the stationary third-order autocorrelation process (AR(3)) given by

$$(1 - \rho_1 L)(1 - \rho_2 L)(1 - \rho_3 L)\varepsilon_t = u_t$$

with $(\rho_1, \rho_2, \rho_3) = (0.05, 0.1, 0.15)$ and $(0.05, 0.1, 0.2)$ and $u_t \sim IN(0, 1)$; and

(iv) the stationary fourth-order autocorrelation process (AR(4)) given by

$$(1 - \rho_1 L)(1 - \rho_2 L)(1 - \rho_3 L)(1 - \rho_4 L)\varepsilon_t = u_t$$

with $(\rho_1, \rho_2, \rho_3, \rho_4) = (0.05, 0.1, 0.15, 0.15)$, $(0.05, 0.1, 0.05, 0.05)$ and $(0.05, 0.05, 0.05, 0.05)$ and $u_t \sim IN(0, 1)$.

All data generating processes were run for $T = 50, 100, 200$ and 500 ; all tests were applied at the 10%, 5% and 1% significance levels and, for the new test procedure, m and n were set to 20,000.

3.1.2 Results

The size results are given in Table 1, and the power results are presented in Table 2. With respect to sizes, all three tests have appropriate sizes and there are no discernible differences with respect to the three tests. The new procedure using either method of bandwidth selection appears to be doing an excellent job of controlling size.

Turning to the power results, an obvious feature is that for the new procedure, the MCMC method of choosing bandwidths does seem to have a slight edge in terms of power over the much simpler NRR method of bandwidth selection.

Overall, the new procedure using MCMC bandwidth selection seems to be the most powerful procedure, although there are a small number of simulations in which the Portmanteau test is most powerful. These tend to be only for $T = 500$ when $d = 4$, but for $d = 6$, it can happen for a larger range of sample sizes. In general, it does appear that the advantage of the new test using MCMC bandwidth selection is greatest for $d = 4$ and declines slightly as we move to the

more complex case of $d = 6$. Also, the new procedure's advantage seems to be greater for lower order alternatives (AR(1), AR(2)) although it should be acknowledged that this may be largely a function of our choice of parameters values.

3.2 Testing for Normality

3.2.1 The Experimental Design

In many statistical situations, random observations are often assumed to be normally distributed for the purpose of statistical inferences. Therefore, it is important to be able to test for normality (see for example, Shapiro and Wilk, 1965; D'Agostino, 1971, 1972; Bowman and Shenton, 1975; Pearson, D'Agostino and Bowman, 1977; Jarque and Bera, 1980, 1987; Spiegelhalter, 1980; Thode, 2002).

The second Monte Carlo experiment involved the problem of testing the null hypothesis that a simple random sample is independently and identically normally distributed with unknown mean (μ) and unknown variance (σ^2) against the alternative that it is non-normally distributed. In other words, (3) is the model for the null hypothesis. Evidence of non-normality is often obtained from sample measures of skewness and kurtosis denoted as $\sqrt{b_1}$ and b_2 , respectively, where

$$b_1 = \hat{\mu}_3^2 / \hat{\mu}_2^3, \quad b_2 = \hat{\mu}_4 / \hat{\mu}_2^2,$$

and $\hat{\mu}_i = \sum_{t=1}^T (y_t - \hat{\mu})^i / T$, for $i = 1, 2, 3, 4$, with $\hat{\mu} = \sum_{t=1}^T y_t / T$. Jarque and Bera (1980, 1987), D'Agostino and Stephens (1986), Urzúa (1996) and Thode (2002) have discussed omnibus tests for normality that combine information from $\sqrt{b_1}$ and b_2 .

This experiment involves comparing the small-sample properties of our test procedure based on the test vector $(\sqrt{b_1}, b_2)'$ with the Jarque-Bera test (Jarque and Bera, 1980, 1987) and the modified version of the normality test proposed by Urzúa (1996). The respective test statistics are

$$\text{JB} = T \left[\frac{(\sqrt{b_1})^2}{6} + \frac{(b_2 - 3)^2}{24} \right], \quad (5)$$

and

$$\text{MJB} = \left[\frac{(\sqrt{b_1})^2}{\text{Var}(\sqrt{b_1})} + \frac{(b_2 - E(b_2))^2}{\text{Var}(b_2)} \right], \quad (6)$$

where $E(b_2) = 3(T - 1)/(T + 1)$, $\text{Var}(\sqrt{b_1}) = 6(T - 2)/[(T + 1)(T + 3)]$ and $\text{Var}(b_2) = 24T(T - 2)(T - 3)/[(T + 1)^2(T + 3)(T + 5)]$.

A small number of simulation studies have revealed that the size of the JB test is incorrect for small- and moderate-sized samples particularly in the context of the linear regression model. The MJB test provides a slight improvement. A more straightforward solution is to use Monte Carlo simulations to obtain correct critical values which is the approach we used in this study (see for example, Dufour and Khalaf, 2001; Poitras, 2006).

Sizes were calculated by simulating (3) with $\varepsilon_t \sim IN(0, 1)$. Powers were calculated for three alternative distributions for ε_t , these being the Stable distribution with dispersion parameter 1.6 and skewness parameter 0 denoted Stable(1.6,0), the Student's t distribution with 5 degrees of freedom denoted t_5 and the Chi-squared distribution with 3 degrees of freedom denoted χ_3^2 . The four tests were compared for sample sizes of $T = 30, 50, 75$, and 100 with the values of m and n both being 20,000.

3.2.2 Results

The size and power results are given in Tables 3 and 4. All four tests have excellent sizes, and just as for testing for autocorrelation, there is no discernible difference between the estimated sizes of the four tests.

With regards to power, the results largely fall into two categories, those for the symmetric distributions (Stable(1.6,0) and t_5) and those for the skewed distribution χ_3^2 . For the first set (Stable(1.6,0) and t_5), there are only small differences in power between the JB and MJB tests with the MJB test being slightly more powerful. There are only small differences in power between the two versions of the new procedure with the test using NRR bandwidths providing a slight advantage, particularly for smaller samples. Almost always, the new procedure has a small power

advantage over both the JB and MJB tests. Only in 3 out of 24 cases considered is the MJB test the most powerful test.

Under the χ_3^2 distribution, the power results show some different patterns. Now the JB test is more powerful than the MJB test. The MCMC based new test is typically more powerful than the NRR based test for smaller sample sizes with that advantage being lost as the sample size increases. Almost always both versions of the new test are more powerful than the JB and MJB tests with some very large improvements in power being evident for smaller sample sizes at the 1% and 5% significance levels. Strangely, the JB test has power being equal to or slightly better than both versions of the new procedure at the 10% significance level.

4 The Testing Procedure for Non-Invariant Test Statistics

So far we have concentrated on test statistics that are invariant to nuisance parameters under the null hypothesis. In this case, there is no issue of how to simulate \mathbf{t} under the null hypothesis in order to estimate its density. When the distribution of \mathbf{t} under the null hypothesis depends on the value of one or more nuisance parameters which we denote by γ , then based on bootstrapping principles, we recommend the following variation to Step (ii) in the procedure given in Section 2.1:

- (ii') Estimate γ assuming the null hypothesis is true and denote this estimate as $\hat{\gamma}$. Using $\gamma = \hat{\gamma}$ and any convenient values of the remaining parameters in the model under the null hypothesis, simulate the model m times and calculate m independent values of \mathbf{t} denoted as $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_m$.

The remainder of the procedure is as outlined in Section 2.1.

5 The Information Matrix Test

It is often important to test whether a model is correctly specified. White (1982) showed that when a model is correctly specified and estimated by maximizing the likelihood function, the information matrix should be asymptotically equal to the negative Hessian matrix. The information matrix test introduced by White (1982), aims to test the significance of the discrepancy between the negative Hessian and the outer product of the score vector, where the lower triangular components of the matrix of such differences are organized into a vector which we call the test vector in this paper. Chesher (1984) showed that the IM test can be viewed as a Lagrange multiplier (LM) test for specification error against the alternative of parameter heterogeneity. Chesher (1983) and Lancaster (1984) presented an nR^2 version of the IM test, where n is the sample size and R^2 is the goodness of fit obtained through the ordinary least squares regression of a column of ones on a matrix whose elements are functions of the first and second derivatives of the log-likelihood function. For the normal fixed regressor linear model, Hall (1987) showed that the LM version of the IM test can be asymptotically decomposed into the sum of three components, where one is White's (1982) general test for heteroscedasticity, and the other two components aim to test certain forms of normality.

The use of the IM test in applied econometrics is limited because the actual size of the test obtained according to the asymptotic critical value often differs greatly from its nominal size. This phenomenon has been evidenced by the Monte Carlo experiments reported in Taylor (1987), Orme (1990), Chesher and Spady (1991) and Davidson and MacKinnon (1992). Davidson and MacKinnon (1992) proposed to deal with this problem by using the double-length artificial regressions to compute a variant of the IM test statistic, but models for discrete, censored or truncated data cannot be dealt with via this method. Chesher and Spady (1991) proposed to obtain the critical value for the IM test from the Edgeworth expansion through order $O(n^{-1})$ of the finite-sample distribution of the test statistic. Their Monte Carlo investigation indicates that such a critical value provides a good approximation to the true critical value obtained through the exact distribution of the IM test, and such an approximation was found to be superior to the usual χ^2 approximation in some cases. In the examples considered by Chesher and Spady (1991),

the Edgeworth expansions are independent of the parameters of the models being tested, but this is not the case in general. Horowitz (1994) proposed a bootstrapping procedure to obtain critical values for the IM test and demonstrated the capability of bootstrapping to overcome the incorrect-size problem in finite samples. Horowitz (1994) showed that in many important circumstances, one can easily obtain good finite-sample critical values for the IM test through bootstrapping rather than through Edgeworth expansions or other algebraically complicated manipulations. Moreover, Horowitz (1994) discussed the power performance of three versions of the IM test through Monte Carlo simulation. His results showed that all three versions of the IM test considered have much lower powers computed according to size-corrected critical values than those computed according to asymptotic critical values. Therefore, it seems that getting the size right and achieving higher power are different tasks.

Most existing versions of the IM test rely on the estimate of the asymptotic covariance matrix of the test vector. The analytical form of the asymptotic covariance of the test vector is complicated and involves the third derivative of the log-likelihood function. Lancaster (1984) showed that the covariance matrix of White's (1982) IM test can be estimated without calculating the third derivative of the log-likelihood. Dhaene and Hoorelbeke (2004) indicated that the incorrect-size problem results from the inaccurate estimate of the covariance matrix of the test vector. They proposed to estimate the covariance matrix of the test vector through parametric bootstrapping.

Let $f(y|\theta)$ denote the density for a postulated model where θ is a $d \times 1$ vector of parameters. Let $\mathbf{y} = (y_1, y_2, \dots, y_T)'$ be the vector of observations, and $\ell(\mathbf{y}|\theta) = \log f(\mathbf{y}|\theta)$ the logarithmic density. We introduce the following notation.

$$A(\theta) = E \left[\frac{\partial^2 \ell(\mathbf{y}|\theta)}{\partial \theta \partial \theta'} \right], \quad A_T(\mathbf{y}, \theta) = \frac{1}{T} \sum_{t=1}^T \frac{\partial^2 \ell(y_t|\theta)}{\partial \theta \partial \theta'},$$

$$B(\theta) = E \left[\frac{\partial \ell(\mathbf{y}|\theta)}{\partial \theta} \frac{\partial \ell(\mathbf{y}|\theta)}{\partial \theta'} \right], \quad B_T(\mathbf{y}, \theta) = \frac{1}{T} \sum_{t=1}^T \frac{\partial \ell(y_t|\theta)}{\partial \theta} \frac{\partial \ell(y_t|\theta)}{\partial \theta'},$$

where expectations are taken with respect to the true density. When the model is correctly specified, the true density is $f(\mathbf{y}|\theta)$. Let θ_0 be the true value of θ .

The information matrix procedure is based on the information-matrix equality, which states

that $A(\theta_0) + B(\theta_0) = 0$ when the model is correctly specified. Given the vector of T independent observations, \mathbf{y} , the information-matrix test investigates the statistical significance of $A_T(\mathbf{y}, \hat{\theta}) + B_T(\mathbf{y}, \hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimator of θ .

Let \mathbf{t} denote the vector of indicators (test vector) whose elements are D_{ij} , for $i = 1, 2, \dots, d$, and $j = 1, \dots, i$, where

$$D_{ij} = \frac{1}{T} \sum_{t=1}^T \left[\frac{\partial \ell(y_t | \hat{\theta})}{\partial \hat{\theta}_i} \frac{\partial \ell(y_t | \hat{\theta})}{\partial \hat{\theta}_j} + \frac{\partial^2 \ell(y_t | \hat{\theta})}{\partial \hat{\theta}_i \partial \hat{\theta}_j} \right]. \quad (7)$$

White (1982) shows that under regularity conditions, the IM test statistic is of the form

$$\xi = n\mathbf{t}'\hat{V}^{-1}\mathbf{t}. \quad (8)$$

where \hat{V} is the consistent estimator of the covariance matrix of \mathbf{t} under H_0 . Under the null hypothesis, ξ is distributed asymptotically χ_q^2 with $q = d(d+1)/2$, and this is based on the asymptotic null distribution of $\mathbf{t} \sim N(0, V(\theta))$. The IM test statistic depends on the estimate of covariance matrix which in turn depends on the estimate of θ . Our proposed testing procedure aims to estimate the joint density of the vector of indicators, \mathbf{t} , and does not depend on \hat{V} .

Since $A_T(\mathbf{y}, \hat{\theta}) + B_T(\mathbf{y}, \hat{\theta})$ is a symmetric matrix, a test of the complete IM identity can be based on the lower triangular elements of $A_T(\mathbf{y}, \hat{\theta}) + B_T(\mathbf{y}, \hat{\theta})$ (or D_{ij}). However, according to White (1982), in many situations it is inappropriate to base the test on all q indicators because some indicators may be identically zero, furthermore, some indicators may be linear combinations of other indicators. In either case, it is appropriate to ignore such indicators. In the remainder of this paper, the IM tests are based on the maximum number of linearly independent non-zero indicators.

6 Simulation Study of the New Testing Procedures Applied to the IM Test

This section reports a Monte Carlo simulation study which aims to compare the sample size power performance of the proposed method with the Lancaster (1984) form of the IM test denoted by

IM_L and the Dhaene and Hoorelbeke (2004) form of the IM test denoted by IM_{DH} . The study covers two different settings. The null hypothesis in the first setting is the normal linear regression model given by

$$y_t = x_t' \beta + u_t, \quad (9)$$

for $t = 1, 2, \dots, T$, where $u_t \sim IN(0, \sigma^2)$, x_t is a $d \times 1$ vector of regressors, and β is a $d \times 1$ vector of parameters. Following Dhaene and Hoorelbeke (2004), we examine the power of the IM test under the heteroscedastic alternative of

$$y_t = x_t' \beta + u_t, \quad u_t \sim N(0, \sqrt{|x_t' \beta|}), \quad \text{for } t = 1, 2, \dots, T. \quad (10)$$

The null model in the second setting is the Tobit model given by

$$y_t = \begin{cases} x_t' \beta + u_t & \text{if } x_t' \beta + u_t > 0 \\ 0 & \text{if } x_t' \beta + u_t \leq 0 \end{cases}, \quad (11)$$

for $t = 1, 2, \dots, T$, $u_t \sim N(0, \sigma^2)$, x_t is a $d \times 1$ vector of regressors, and β is a $d \times 1$ vector of parameters. It will be convenient to re-parameterize the model as

$$hy_t = \begin{cases} x_t' b + v_t & \text{if } x_t' b + v_t > 0 \\ 0 & \text{if } x_t' b + v_t \leq 0 \end{cases}, \quad (12)$$

where $h = 1/\sigma$, $b = \beta/\sigma$, and $v_t \sim N(0, 1)$, for $t = 1, 2, \dots, T$.

Following Horowitz (1994), we examine the power of the IM tests under the models given by

$$y_t = \max(0, x_t' \beta + u_t), \quad u_t \sim N\left(0, \sqrt{\exp(0.5 x_t' \beta)}\right), \quad (13)$$

and

$$y_t = \max(0, x_t' \beta + 0.75x_{t,2}x_{t,3} + u_t), \quad u_t \sim N(0, 1), \quad (14)$$

for $t = 1, 2, \dots, T$, where $x_{t,2}$ and $x_{t,3}$ are the two non-intercept components of x_t . Note that model (13) involves a heteroscedastic alternative while model (14) has an incorrect mean function.

For the linear regression model (9), the IM test statistic is pivotal and the proposed method of testing is therefore based on the procedure outlined in Section 2. In the case of the Tobit model,

the IM test statistic is not pivotal under the null hypothesis and therefore the proposed testing procedure is based on the bootstrapping approach outlined in Section 4.

The experiments consist of applying both forms of IM tests along with the proposed method of testing to the linear regression model and the Tobit model. In both models, x_t , a vector of explanatory variables, consists of an intercept component and either one or two additional variables. The values of x_t are fixed in repeated samples. The values of the β parameters are $(0.75, 1)'$ when x_t consists of an intercept and one regressor, and $(0.75, 1, 1)'$ when x_t consists of an intercept and two regressors. The non-intercept components of x_t are sampled independently either from the standard normal distribution or from the uniform distribution on $(-1, 1)$. The value of σ^2 is 1 in all of the experiments. The sample sizes are 50, 100, 200 and 300. Size-corrected critical values, which were obtained via simulation under the null hypothesis with known true parameters, were used for computing the sizes and powers of the IM_L and IM_{DH} tests. For the IM_{DH} test statistic, we used 50 parametric bootstrap samples to estimate the covariance matrix, \hat{V} , following Dhaene and Hoorelbeke (2004). It should be noted that when the IM test statistic is not pivotal (i.e. for the Tobit model), these size-corrected critical values for the IM_L and IM_{DH} tests cannot be calculated in a practical application because the true parameter values under the null hypothesis are unknown. We have used these critical values in the simulation so that the powers of the respective tests can be compared fairly.

7 Simulation Results

7.1 Results from the Linear Model

The results for the linear model and the case where x_t are sampled independently from standard normal distribution presented in Tables 5 and 6 for sizes and powers, respectively. From Table 5, we see that the sizes derived through the proposed test and both versions of IM test are very close to their corresponding nominal sizes for both one-regressor and two-regressors models. The sizes obtained through all methods appear not to be significantly different from the corresponding

nominal sizes. From Table 6, we see that the power of the proposed method of testing is always higher and often vastly higher than both IM_L and IM_{DH} tests for all sample sizes and nominal sizes. In term of accuracy, the proposed method of testing not only can produce correctly estimated sizes but also has much higher powers than the both versions of IM tests.

We obtained similar results when x_i are sampled independently from uniform distribution on $(-1,1)$, and in the interest of space, these results are not presented here.

7.2 Results from the Tobit Model

The size and power results for the Tobit model are presented in Tables 7, 8 and 9. In Table 7, we see that the sizes derived through the proposed testing procedure are very close to the corresponding nominal sizes for both one-regressor and two-regressors models, and whether the regressor vector x_t is generated through standard normal distribution or uniform distribution. On the other hand, the sizes for IM_L and IM_{DH} tests have mixed behavior. At the 1% level, the sizes seem to be over rejecting the null hypothesis, while at 5% and 10%, levels, the sizes are close to their nominal sizes. This behavior is consistent for both regressor vectors as well as for both one-regressor and two-regressor models. However, the sizes derived through all tests do not appear to be significantly different from their corresponding nominal sizes.

Table 8 presents the estimated powers of the tests, where model (13) is used as a true alternative hypothesis. We found that the proposed test has higher powers than both versions of the IM test. This is consistent for both one-regressor and two-regressor models, and whether the regressor vector x_i is generated through the standard normal distribution or the uniform distribution. Moreover, the powers obtained through one-regressor model are smaller than those derived through the two-regressors model for almost all sample sizes and all nominal sizes. This is likely to be because the two-regressor model has a higher degree of heteroscedasticity. Thus, the simulation study shows that the proposed test produce correct sizes and has higher powers than the Lancaster and DH versions of IM test.

Table 9 presents² the estimated powers of the IM_L and our proposed test when model (14) is the true model. Again we see our proposed test almost always having a considerable power advantage over the IM_L test — the only exceptions being for smaller sample sizes and $\alpha = 0.01$ in the two-regressor case which may be caused by the size differences of the two tests in this setting.

8 Conclusion

This paper presents a new testing procedure, in which we estimate the density of the test vector whose components are the multiple test statistics through simulation and kernel density estimation rather than constructing a critical region through a scalar test statistic. Using the estimated density, we are then able to approximate the overall p value of the multiple test statistics. In the case where the distribution of the test statistics depend on nuisance parameters under the null hypothesis, they are first estimated, and then the model is simulated for these estimated values of the nuisance parameters to allow kernel density estimation to take place.

In order to examine the size and power of the proposed testing procedure, we have conducted a two-stage procedure of Monte Carlo simulations, where the first-stage simulation aims to estimate the density of the test vector, while the second-stage simulation involves estimating the size and powers the testing procedure by the relative frequency of rejecting the null hypothesis. Our testing procedure has been compared with respectively, the Portmanteau test for testing autocorrelations, two versions of Jarque and Bera’s (1970) test for normality, and two versions of White’s (1982) information matrix test for model misspecification. The simulation studies have shown that our testing procedure has correct or nearly correct sizes, and that the power of our testing procedure is better than, or in some cases as good as, any of the competing tests.

It appears that for relatively simple testing problems with few nuisance parameters, such as testing for autocorrelation and non-normality in a random sample, the new procedure typically has a slight advantage in terms of power. We see evidence of that advantage increasing as we turn

²We expect to have the corresponding power results for the IM_{DH} test when this paper is next revised. Preliminary calculations suggest that the IM_L and IM_{DH} tests have very similar powers in this case.

to the more difficult problem of testing for misspecification via the information matrix in the linear regression model and the Tobit model. The standard approach in this more complicated setting is to derive the asymptotic distribution of the vector of statistics, estimate the asymptotic covariance matrix and calculate the usual quadratic form that has an asymptotic Chi-squared distribution under the null hypothesis. Each of the steps involves an approximation that has the potential to affect the power of the resultant test. Our approach focuses directly on the small-sample null distribution of the vector of statistics in order to estimate the overall p value. We believe it is the first good method for calculating the overall p value for a vector of test statistics, based on simulation.

An important step in our procedure is the selection of bandwidth values for kernel density estimation. We found that the MCMC based approach is slightly better than the NRR although there is a very big difference in the computational time required. So if this is an issue then the use of the NRR can provide very acceptable results.

Acknowledgements

We gratefully acknowledge computational support from the Victorian Partnership for Advanced Computing (VPAC). This research was supported under Australian Research Council's *Discovery Projects* funding scheme.

References

- Bowman, A.W., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis. Oxford University Press, London.
- Bowman, K.O., Shenton, L.R., 1975. Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2 . *Biometrika* 62, 243-250.
- 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.

- Chesher, A., 1983. The information matrix test: Simplified calculation via a score test interpretation. *Economic letters* 13, 45-48.
- Chesher, A., 1984. Testing for neglected heterogeneity. *Econometrica* 52, 865-872.
- Chesher, A., Spady, R., 1991. Asymptotic expansions of the information matrix test statistic. *Econometrica* 59, 787-815.
- D'Agostino, R.B., 1971. An omnibus test of normality for moderate and large size samples. *Biometrika* 58, 341-348.
- D'Agostino, R.B., 1972. Small sample probability points for the D test of normality. *Biometrika* 59, 219-221.
- D'Agostino, R.B., Stephens, M., 1986. *Goodness-of-fit Techniques*. Marcel Dekker, New York.
- Davidson, R., MacKinnon, J.G., 1992. A new form of the information matrix test. *Econometrica* 60, 145-157.
- Dhaene, G., Hoorelbeke, D., 2004. The information matrix test with bootstrap-based covariance matrix estimation. *Economics Letters* 82, 341-347.
- Dufour, J., Khalaf, L., 2001. Monte Carlo test methods in econometrics. In Baltagi, B. (Eds.), *Companion to Theoretical Econometrics*. Blackwell, Oxford.
- Durfour J., Farhat, A., Gardial, L., Khalaf, L., 1998. Simulation-based finite sample normality tests in linear regressions. *Econometrics Journal* 1, C154-C173.
- Hall, A., 1987. The information matrix test for the linear model. *Review of Economic Studies* 54, 257-263.
- Hyndman, R.J., 1996. Computing and graphing highest density region. *The American Statistician* 50, 120-126.
- Horowitz, J.L., 1994. Bootstrap-based critical values for the information matrix test. *Journal of Econometrics* 61, 395-411.
- Jarque, C.M., Bera, A.K., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* 6, 255-259.

- Jarque, C.M., Bera, A.K., 1987. A Test for normality of observations and regression residuals. *International Statistical Review* 55, 163-172.
- Lancaster, T., 1984. The covariance matrix of the information matrix test. *Econometrica* 52, 1051-1053.
- Ljung, L., Box, G.E.P., 1987. On a measure of lack of fit in time series models. *Biometrika* 66, 67-72.
- Orme, C., 1990. The small-sample performance of the information matrix test. *Journal of Econometrics* 46, 309-331.
- Pearson, E.S., D'Agostino, R.B., Bowman, K.O., 1977. Test for departure from normality: comparison of powers. *Biometrika* 64, 231-246.
- Poitras, G., 1992. Testing regression disturbances for normality with stable alternatives: further Monte Carlo evidence. *Journal of Statistical Computation and Simulation* 41, 109-123.
- Poitras, G., 2006. More on the correct use of omnibus tests for normality. *Economics Letters* 90, 304-309.
- Scott, D.W., 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York.
- Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test of normality (complete samples). *Biometrika* 52, 591-611.
- Spiegelhalter, D.J., 1980. An omnibus test for normality for small samples. *Biometrika* 67, 493-496.
- Taylor, L.W., 1987. The size bias of White's information matrix test. *Economics Letters* 24, 63-68.
- Thode, H., 2002. *Testing for Normality*. Marcel Dekker, New York.
- Urzúa, C.M., 1996. On the correct use of omnibus tests for normality. *Economics Letters* 53, 247-251.
- Wand, M.P., Jones, M.C., 1995. *Kernel Smoothing*. Chapman & Hall, London.
- White, H., 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1-26.

Zhang, X., King, M.L., Hyndman, R.J., 2006. A Bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics and Data Analysis* 50, 3009-3031.

Table 1: *The estimated sizes of the new test procedures and the Portmanteau test when $d = 4$ and $d = 6$.*

Dimension	T	New test (NRR)			New test (MCMC)			Portmanteau		
		0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
$d = 4$	50	0.010	0.051	0.100	0.010	0.051	0.099	0.010	0.052	0.100
	100	0.009	0.048	0.098	0.009	0.049	0.101	0.011	0.051	0.102
	200	0.010	0.050	0.101	0.011	0.050	0.102	0.010	0.052	0.102
	500	0.009	0.045	0.097	0.009	0.046	0.095	0.009	0.045	0.095
$d = 6$	50	0.010	0.054	0.104	0.010	0.054	0.103	0.010	0.053	0.104
	100	0.010	0.049	0.099	0.010	0.050	0.100	0.011	0.050	0.103
	200	0.012	0.052	0.101	0.011	0.052	0.100	0.010	0.051	0.103
	500	0.008	0.044	0.092	0.008	0.045	0.092	0.009	0.046	0.094

Table 2: *Estimated powers of the new test procedures and the Portmanteau test when $d = 4$.*

Alternative hypothesis	Coefficients	T	New test (NRR)			New test (MCMC)			Portmanteau		
			0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
$AR(2)$	$\rho_1 = 0.05$ $\rho_2 = 0.10$	50	0.023	0.093	0.162	0.025	0.093	0.163	0.019	0.082	0.139
		100	0.042	0.133	0.218	0.044	0.136	0.223	0.037	0.120	0.201
		200	0.082	0.219	0.324	0.085	0.222	0.330	0.082	0.212	0.312
		500	0.262	0.494	0.621	0.274	0.500	0.624	0.278	0.492	0.615
$AR(2)$	$\rho_1 = 0.05$ $\rho_2 = 0.20$	50	0.078	0.195	0.296	0.079	0.199	0.301	0.054	0.167	0.254
		100	0.171	0.360	0.486	0.176	0.365	0.493	0.153	0.326	0.448
		200	0.402	0.645	0.755	0.417	0.651	0.760	0.392	0.624	0.733
		500	0.901	0.975	0.987	0.913	0.976	0.988	0.911	0.971	0.987
$AR(3)$	$\rho_1 = 0.10$ $\rho_2 = 0.10$ $\rho_3 = 0.10$	50	0.062	0.162	0.250	0.064	0.162	0.253	0.048	0.138	0.211
		100	0.141	0.295	0.404	0.145	0.301	0.410	0.145	0.284	0.382
		200	0.316	0.535	0.646	0.334	0.542	0.654	0.356	0.551	0.651
		500	0.781	0.919	0.955	0.802	0.925	0.957	0.843	0.932	0.960
$AR(3)$	$\rho_1 = 0.05$ $\rho_2 = 0.05$ $\rho_3 = 0.20$	50	0.071	0.187	0.284	0.071	0.187	0.283	0.047	0.154	0.234
		100	0.177	0.372	0.490	0.183	0.373	0.493	0.153	0.327	0.446
		200	0.422	0.678	0.780	0.441	0.684	0.786	0.417	0.653	0.758
		500	0.926	0.980	0.991	0.933	0.980	0.991	0.932	0.978	0.991
$AR(4)$	$\rho_1 = 0.05$ $\rho_2 = 0.10$ $\rho_3 = 0.15$ $\rho_4 = 0.15$	50	0.137	0.273	0.380	0.138	0.274	0.382	0.102	0.227	0.306
		100	0.331	0.538	0.647	0.344	0.544	0.652	0.316	0.492	0.594
		200	0.676	0.846	0.902	0.699	0.851	0.905	0.684	0.830	0.887
		500	0.987	0.997	0.999	0.990	0.998	0.999	0.992	0.997	0.999
$AR(4)$	$\rho_1 = 0.05$ $\rho_2 = 0.10$ $\rho_3 = 0.05$ $\rho_4 = 0.10$	50	0.058	0.154	0.235	0.058	0.154	0.237	0.043	0.125	0.191
		100	0.116	0.267	0.378	0.121	0.272	0.383	0.114	0.241	0.334
		200	0.248	0.470	0.592	0.265	0.480	0.602	0.277	0.467	0.572
		500	0.698	0.868	0.924	0.718	0.877	0.925	0.743	0.871	0.919
$AR(4)$	$\rho_1 = 0.05$ $\rho_2 = 0.05$ $\rho_3 = 0.05$ $\rho_4 = 0.05$	50	0.023	0.087	0.152	0.024	0.090	0.155	0.018	0.072	0.126
		100	0.039	0.126	0.209	0.040	0.128	0.210	0.038	0.111	0.183
		200	0.078	0.205	0.303	0.082	0.208	0.308	0.086	0.200	0.286
		500	0.127	0.292	0.406	0.234	0.457	0.576	0.135	0.287	0.387

Table 3: *Estimated powers of the new test procedures and the Portmanteau test when $d = 6$.*

Alternative hypothesis	Coefficients	T	New test (NRR)			New test (MCMC)			Portmanteau		
			0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
$AR(2)$	$\rho_1 = 0.05$ $\rho_2 = 0.10$	50	0.007	0.056	0.125	0.009	0.065	0.133	0.017	0.076	0.134
		100	0.023	0.109	0.189	0.025	0.113	0.197	0.031	0.105	0.180
		200	0.070	0.187	0.280	0.070	0.195	0.287	0.065	0.187	0.278
		500	0.211	0.421	0.550	0.218	0.438	0.568	0.219	0.430	0.559
$AR(2)$	$\rho_1 = 0.05$ $\rho_2 = 0.20$	50	0.052	0.167	0.260	0.055	0.172	0.264	0.044	0.146	0.227
		100	0.128	0.301	0.418	0.134	0.316	0.433	0.117	0.274	0.394
		200	0.337	0.562	0.680	0.345	0.583	0.696	0.318	0.560	0.675
		500	0.850	0.947	0.974	0.863	0.957	0.979	0.862	0.955	0.976
$AR(3)$	$\rho_1 = 0.10$ $\rho_2 = 0.10$ $\rho_3 = 0.10$	50	0.048	0.137	0.217	0.051	0.143	0.221	0.040	0.120	0.188
		100	0.112	0.253	0.355	0.116	0.265	0.364	0.115	0.244	0.341
		200	0.271	0.469	0.587	0.277	0.487	0.600	0.296	0.497	0.602
		500	0.722	0.872	0.927	0.739	0.890	0.938	0.790	0.906	0.943
$AR(3)$	$\rho_1 = 0.05$ $\rho_2 = 0.05$ $\rho_3 = 0.20$	50	0.056	0.159	0.247	0.057	0.164	0.249	0.039	0.130	0.211
		100	0.141	0.311	0.429	0.141	0.317	0.436	0.118	0.276	0.391
		200	0.362	0.588	0.704	0.373	0.608	0.723	0.346	0.588	0.702
		500	0.884	0.959	0.981	0.892	0.966	0.983	0.894	0.965	0.982
$AR(4)$	$\rho_1 = 0.05$ $\rho_2 = 0.10$ $\rho_3 = 0.15$ $\rho_4 = 0.15$	50	0.104	0.236	0.333	0.106	0.240	0.336	0.081	0.194	0.275
		100	0.272	0.469	0.577	0.277	0.482	0.591	0.268	0.437	0.540
		200	0.609	0.789	0.862	0.622	0.809	0.872	0.624	0.790	0.852
		500	0.976	0.994	0.997	0.980	0.996	0.998	0.985	0.996	0.998
$AR(4)$	$\rho_1 = 0.05$ $\rho_2 = 0.10$ $\rho_3 = 0.05$ $\rho_4 = 0.10$	50	0.041	0.126	0.202	0.042	0.129	0.206	0.034	0.108	0.171
		100	0.090	0.224	0.329	0.091	0.232	0.338	0.088	0.205	0.298
		200	0.208	0.405	0.530	0.212	0.423	0.545	0.230	0.412	0.522
		500	0.630	0.809	0.881	0.647	0.829	0.894	0.683	0.833	0.891
$AR(4)$	$\rho_1 = 0.05$ $\rho_2 = 0.05$ $\rho_3 = 0.05$ $\rho_4 = 0.05$	50	0.013	0.056	0.102	0.014	0.061	0.112	0.015	0.067	0.120
		100	0.026	0.098	0.173	0.027	0.104	0.182	0.029	0.096	0.166
		200	0.066	0.173	0.266	0.066	0.182	0.273	0.068	0.175	0.260
		500	0.190	0.376	0.503	0.098	0.395	0.523	0.219	0.405	0.519

Table 4: *The estimated sizes and powers of the new test procedures, JB and MJB tests*

Size and power	T	New test (NRR)			New test (MCMC)			JB			MJB		
		0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
Size	30	0.010	0.049	0.098	0.010	0.048	0.099	0.010	0.049	0.096	0.009	0.048	0.097
	50	0.008	0.047	0.095	0.008	0.046	0.094	0.009	0.045	0.095	0.008	0.047	0.098
	75	0.008	0.049	0.097	0.008	0.048	0.097	0.008	0.049	0.097	0.008	0.050	0.097
	100	0.009	0.045	0.096	0.008	0.046	0.095	0.009	0.047	0.098	0.009	0.048	0.097
Power Stable(1.6, 0)	30	0.468	0.599	0.671	0.468	0.598	0.671	0.464	0.587	0.651	0.467	0.596	0.665
	50	0.670	0.778	0.824	0.670	0.778	0.824	0.665	0.764	0.809	0.667	0.771	0.819
	75	0.815	0.889	0.918	0.817	0.890	0.918	0.811	0.884	0.907	0.812	0.888	0.913
	100	0.899	0.945	0.960	0.899	0.945	0.960	0.896	0.942	0.957	0.897	0.944	0.960
t_5	30	0.175	0.323	0.425	0.174	0.322	0.423	0.176	0.314	0.396	0.178	0.319	0.414
	50	0.276	0.451	0.543	0.276	0.450	0.542	0.273	0.429	0.523	0.274	0.442	0.540
	75	0.370	0.564	0.659	0.372	0.564	0.661	0.370	0.548	0.629	0.370	0.559	0.648
	100	0.470	0.650	0.734	0.468	0.650	0.734	0.461	0.640	0.715	0.465	0.651	0.728
χ_3^2	30	0.472	0.742	0.840	0.486	0.754	0.846	0.391	0.688	0.850	0.359	0.629	0.790
	50	0.821	0.952	0.979	0.824	0.953	0.979	0.634	0.905	0.980	0.595	0.872	0.963
	75	0.976	0.997	0.999	0.974	0.996	0.999	0.840	0.990	0.999	0.804	0.982	0.998
	100	0.998	1.000	1.000	0.998	1.000	1.000	0.951	0.999	1.000	0.935	0.999	1.000

Table 5: *Estimated sizes of the IM tests with samples generated from the linear regression model*

Test	Sample size	One-regressor			Two-regressor		
		0.01	0.05	0.10	0.01	0.05	0.10
New test	50	0.010	0.051	0.100	0.011	0.048	0.098
	100	0.009	0.046	0.095	0.010	0.050	0.099
	200	0.009	0.051	0.097	0.008	0.049	0.099
	300	0.010	0.050	0.102	0.009	0.048	0.097
IM_L	50	0.011	0.052	0.103	0.010	0.051	0.102
	100	0.011	0.050	0.100	0.011	0.051	0.100
	200	0.008	0.051	0.103	0.011	0.052	0.101
	300	0.011	0.050	0.103	0.009	0.048	0.103
IM_{DH}	50	0.010	0.053	0.102	0.011	0.051	0.102
	100	0.011	0.049	0.097	0.010	0.049	0.098
	200	0.010	0.051	0.101	0.009	0.050	0.101
	300	0.010	0.054	0.107	0.010	0.048	0.096

Table 6: *Estimated powers of the IM tests with samples generated from the linear model*

Tests	Sample size	One-regressor			Two-regressor		
		0.01	0.05	0.10	0.01	0.05	0.10
New test	50	0.106	0.363	0.533	0.030	0.121	0.216
	100	0.491	0.801	0.913	0.243	0.572	0.743
	200	0.975	0.998	1.000	0.738	0.944	0.976
	300	0.999	1.000	1.000	0.927	0.991	0.998
IM_L	50	0.017	0.074	0.154	0.014	0.055	0.107
	100	0.034	0.268	0.475	0.022	0.130	0.263
	200	0.422	0.803	0.914	0.202	0.500	0.680
	300	0.751	0.962	0.993	0.396	0.781	0.911
IM_{DH}	50	0.025	0.133	0.281	0.020	0.104	0.219
	100	0.078	0.473	0.713	0.051	0.287	0.500
	200	0.580	0.948	0.986	0.196	0.649	0.832
	300	0.910	0.993	0.999	0.485	0.873	0.952

Table 7: *Estimated sizes of the IM tests with samples generated from the Tobit model*

Tests	Sample size	One-regressor			Two-regressor		
		0.01	0.05	0.10	0.01	0.05	0.10
New test	50	0.011	0.055	0.109	0.009	0.049	0.104
	100	0.009	0.048	0.097	0.009	0.055	0.112
	200	0.011	0.057	0.106	0.012	0.046	0.098
	300	0.013	0.047	0.087	0.011	0.054	0.107
IM_L	50	0.018	0.058	0.105	0.012	0.047	0.089
	100	0.016	0.056	0.108	0.015	0.051	0.092
	200	0.016	0.052	0.100	0.013	0.049	0.098
	300	0.014	0.049	0.100	0.017	0.065	0.124
IM_{DH}	50	0.020	0.052	0.102	0.016	0.056	0.100
	100	0.020	0.060	0.100	0.019	0.053	0.107
	200	0.016	0.055	0.112	0.010	0.045	0.116
	300	0.015	0.057	0.111	0.015	0.054	0.107

Table 8: *Estimated powers of the IM tests with samples generated from (13)*

Test	Sample size	One-regressor			Two-regressor		
		0.01	0.05	0.10	0.01	0.05	0.10
New test	50	0.156	0.362	0.503	0.120	0.330	0.471
	100	0.247	0.501	0.629	0.259	0.598	0.726
	200	0.454	0.782	0.866	0.647	0.926	0.951
	300	0.699	0.917	0.948	0.818	0.992	1.000
IM_L	50	0.054	0.128	0.197	0.052	0.103	0.187
	100	0.062	0.198	0.320	0.120	0.234	0.351
	200	0.196	0.425	0.572	0.311	0.652	0.766
	300	0.396	0.674	0.782	0.706	0.841	0.952
IM_{DH}	50	0.043	0.167	0.307	0.029	0.110	0.223
	100	0.112	0.349	0.536	0.097	0.315	0.506
	200	0.365	0.712	0.858	0.475	0.795	0.930
	300	0.622	0.891	0.943	0.762	0.976	1.000

Table 9: *Estimated powers of the IM tests with samples generated from (14)*

Test	Sample size	Significance level		
		0.01	0.05	0.10
New test	50	0.025	0.115	0.201
	100	0.041	0.184	0.320
	200	0.222	0.580	0.742
	300	0.476	0.844	0.930
IM_L	50	0.033	0.077	0.144
	100	0.056	0.117	0.182
	200	0.139	0.291	0.395
	300	0.258	0.440	0.578