Chapter 13

WALD, LIKELIHOOD RATIO, AND LAGRANGE MULTIPLIER TESTS IN ECONOMETRICS

ROBERT F. ENGLE*

University of California

Contents

*Research supported by NSF SOC 78-09476 and 80-08580.

Handbook of Econometrics, Volume II, Edited by Z. Griliches and M.D. Intriligator
© Elsevier Science Publishers BV, 1984

1. Introduction

If the confrontation of economic theories with observable phenomena is the objective of empirical research, then hypothesis testing is the primary tool of analysis. To receive empirical verification, all theories must eventually be reduced to a testable hypothesis. In the past several decades, least squares based tests have functioned admirably for this purpose. More recently, the use of increasingly complex statistical models has led to heavy reliance on maximum likelihood methods for both estimation and testing. In such a setting only asymptotic properties can be expected for estimators or tests. Often there are asymptotically equivalent procedures which differ substantially in computational difficulty and finite sample performance. Econometricians have responded enthusiastically to this research challenge by devising a wide variety of tests for these complex models.

Most of the tests used are based either on the Wald, Likelihood Ratio or Lagrange Multiplier principle. These three general principles have a certain symmetry which has revolutionized the teaching of hypothesis tests and the development of new procedures. Essentially, the Lagrange Multiplier approach starts at the null and asks whether movement toward the alternative would be an improvement, while the Wald approach starts at the alternative and considers movement toward the null. The Likelihood ratio method compares the two hypotheses directly on an equal basis. This chapter provides a unified development of the three principles beginning with the likelihood functions. The properties of the tests and the relations between them are developed and their forms in a variety of common testing situations are explained. Because the Wald and Likelihood Ratio tests are relatively well known in econometrics, major emphasis will be put upon the cases where Lagrange Multiplier tests are particularly attractive. At the conclusion of the chapter, three other principles will be compared: Neyman's (1959) $C(\alpha)$ test, Durbin's (1970) test procedure, and Hausman's (1978) specification test.

2. Definitions and intuitions

Hypothesis testing concerns the question of whether data appear to favor or disfavor a particular description of nature. Testing is inherently concerned with one particular hypothesis which will be called the *null* hypothesis. If the data fall into a particular region of the sample space called the *critical region* then the test is said to *reject* the null hypothesis, otherwise it *accepts*. As there are only two possible outcomes, an hypothesis testing problem is inherently much simpler than an estimation problem where there are a continuum of possible outcomes. It is important to notice that both of these outcomes refer only to the null hypothesis —we either reject or accept it. To be even more careful in terminology, we either reject or fail to reject the null hypothesis. This makes it clear that the data may not contain evidence against the null simply because they contain very little information at all concerning the question being asked.

As there are only two possible outcomes, there are only two ways to make incorrect inferences. *Type I* errors are committed when the null hypothesis is falsely rejected, and *Type II* errors occur when it is incorrectly accepted. For any test we call α the *size* of the test which is the probability of Type I errors and β is the probability of Type II errors. The *power* of a test is the probability of rejecting the null when it is false, which is therefore $1-\beta$.

In comparing tests, the standard notion of optimality is based upon the size and power. Within a class of tests, one is said to be *best* if it has the maximum power (minimum probability of Type II error) among all tests with size (probability of Type I error) less than or equal to some particular level.

To make such conditions operational, it is necessary to specify how the data are generated when the null hypothesis is false. This is the *alternative* hypothesis and it is through careful choice of this alternative that tests take on the behavior desired by the investigator. By specifying an alternative, the critical region can be tailored to look for deviations from the null in the direction of the alternative. It should be emphasized here that rejection of the null does not require accepting the alternative. In particular, suppose some third hypothesis is the true one. It may be that the test would still have some power to reject the null even though it was not the optimal test against the hypothesis actually operating. Another case in point might be where the data would reject the null hypothesis as being implausible, but the alternative could be even more unlikely.

As an example of the role of the alternative, consider the diagnostic problem which is discussed later in Section 7. The null hypothesis is that the model is correctly specified while the alternative is a particular type of problem such as serial correlation. In this case, rejection of the model does not mean that a serial correlation correction is the proper solution. There may be an omitted variable or incorrect functional form which is responsible for the rejection. Thus the serial correlation test has some power against omitted variables even though it is not the optimal test against that particular alternative.

To make these notions more precise and set the stage for large sample results, let y be a $T\times 1$ random vector drawn from the joint density $f(y,\theta)$ where θ is a $k\times 1$ vector of unknown parameters and $\theta\in\Theta$, the parameter space. Under the null $\theta\in\Theta_0\subset\Theta$ and under the alternative $\theta\in\Theta_1\in\Theta$ with $\Theta_0\cap\Theta_1=\emptyset$. Frequently $\Theta_1=\Theta-\Theta_0$. Then for a critical region C_T, the size α_T is given by:

$$\alpha_T=\Pr(y\in C_T|\theta\in\Theta_0). \tag{1}$$

The power of the test is:

$$\pi_T(\theta) = \Pr(y \in C_T|\theta), \quad \text{for } \theta \in \Theta_1. \tag{2}$$

Notice that although the power will generally depend upon the unknown parameter θ, the size usually does not. In most problems where the null hypothesis is *composite* (includes more than one possible value of θ) the class of tests is restricted to those where the size does not depend upon the particular value of $\theta \in \Theta_0$. Such tests are called *similar* tests.

Frequently, there are no tests whose size is calculable exactly or whose size is independent of the point chosen within the null parameter space. In these cases, the investigator may resort to asymptotic criteria of optimality for tests. Such an approach may produce tests which have good finite sample properties and in fact, if there exist exact tests, the asymptotic approach will generally produce them. Let C_T be a sequence of critical regions perhaps defined by a sequence of vectors of statistics $s_T(y) \geq c_T$, where c_T is a sequence of constant vectors. Then the limiting size and power of the test are simply

$$\alpha = \lim_{T \to \infty} \alpha_T; \qquad \pi(\theta) = \lim_{T \to \infty} \pi_T(\theta), \quad \text{for } \theta \in \Theta_1. \tag{3}$$

A test is called *consistent* if $\pi(\theta) = 1$ for all $\theta \in \Theta_1$. That is, a consistent test will always reject the null when it is false; Type II errors are eliminated for large samples if a test is consistent.

As most hypothesis tests are consistent, it remains important to choose among them. This is done by examining the rate at which the power function approaches its limiting value. The most common limiting argument is to consider the power of the test to distinguish alternatives which are very close to the null. As the sample grows, alternatives ever closer to the null can be detected by the test. The power against such *local* alternatives for tests of fixed asymptotic size provides the major criterion for the optimality of asymptotic tests.

The vast majority of all testing problems in econometrics can be formulated in terms of a partition of the parameter space into two sub-vectors $\theta = (\theta_1', \theta_2')'$ where the null hypothesis specifies values, θ_1^0 for θ_1, but leaves θ_2 unconstrained. In a normal testing problem, θ_1 might be the mean and θ_2 the variance, or in a regression context, θ_1 might be several of the parameters while θ_2 includes the rest, the variance and the serial correlation coefficient, if the model has been estimated by Cochrane–Orcutt. Thus θ_1 includes the parameters of interest in the test.

In this context, the null hypothesis is simply:

$$H_0\colon \theta_1 = \theta_1^0, \qquad \theta_2 \text{ unrestricted.} \tag{4}$$

A sequence of local alternatives can be formulated as:

$$H_1\colon \theta_1^T = \theta_1^0 + \delta/T^{1/2}, \qquad \theta_2 \text{ unrestricted}, \tag{5}$$

for some vector δ. Although this alternative is obviously rather peculiar, it serves to focus attention on the portion of the power curve which is most sensitive to the quality of the test. The choice of δ determines in what direction the test will seek departures from the null hypothesis. Frequently, the investigator will chose a test which is equally good in all directions δ, called an *invariant* test.

It is in this context that the optimality of the likelihood ratio test can be established as is done in Section 6. It is asymptotically locally most powerful among all invariant tests. Frequently in this chapter the term *asymptotically optimal* will be used to refer to this characterization. Any tests which have the property that asymptotically they always agree if the data are generated by the null or by a local alternative, will be termed *asymptotically* equivalent. Two tests ξ_1 and ξ_2 with the same critical values will be asymptotically equivalent if $\operatorname{plim}|\xi_1 - \xi_2| = 0$ for the null and local alternatives.

Frequently in testing problems non-linear hypotheses such as $g(\theta) = 0$ are considered where g is a $p \times 1$ vector of functions defined on Θ. Letting the true value of θ under the null be θ^0, then $g(\theta^0) = 0$. Assuming g has continuous first derivatives, expand this in a Taylor series:

$$g(\theta) = g(\theta^0) + G(\bar{\theta})(\theta - \theta^0),$$

where $\bar{\theta}$ lies between θ and θ^0 and $G(\cdot)$ is the first derivative matrix of g. For the null and local alternatives, θ approaches θ^0 so $G(\bar{\theta}) \to G(\theta^0) \equiv G$ and the restriction is simply this linear hypothesis:

$$G\theta = G\theta^0.$$

For any linear hypothesis one can always reparameterize by a linear non-singular matrix $A^{-1}\theta = \phi$ such that this null is H_0: $\phi_1 = \phi_1^0, \phi_2$ unrestricted. To do this let A_2 have $K - p$ columns in the orthogonal complement of G so that $GA_2 = 0$. The remaining p columns of A say A_1, span the row space of G so that GA is non-singular. Then the null becomes:

$$G\theta^0 = G\theta = GA\phi = GA_1\phi_1 + GA_2\phi_2 = GA_1\phi_1,$$

or $\phi_1 = \phi_1^0$ with $\phi_1^0 = (GA_1)^{-1}G\theta^0$.

Thus, for local alternatives there is no loss of generality in considering only linear hypotheses, and in particular, hypotheses which have preassigned values for a subset of the parameter vector.

3. A general formulation of Wald, Likelihood Ratio, and Lagrange Multiplier tests

In this section the basic forms of the three tests will be given and interpreted. Most of this material is familiar in the econometrics literature in Breusch and Pagan (1980) or Savin (1976) and Berndt and Savin (1977). Some new results and intuitions will be offered. Throughout it will be assumed that the likelihood function satisfies standard regularity conditions which allow two term Taylor series expansions and the interchange of integral and derivative. In addition, it will be assumed that the information matrix is non-singular, so that the parameters are (locally) identified.

The simplest testing problem assumes that the data y are generated by a joint density function $f(y,\theta^0)$ under the null hypothesis and by $f(y,\theta)$ with $\theta \in R^k$ under the alternative. This is a test of a simple null against a composite alternative. The log-likelihood is defined as:

$$L(\theta, y) = \log f(y,\theta), \tag{6}$$

which is maximized at a value $\hat{\theta}$ satisfying:

$$\frac{\partial L}{\partial \theta}(\hat{\theta}, y) = 0.$$

Defining $s(\theta, y) = \partial L(\theta, y)/\partial\theta$ as the score, the MLE sets the score to zero. The variance of $\hat{\theta}$ is easily calculated as the inverse of Fisher's Information, or

$$\begin{aligned} V(\hat{\theta}) &= \mathscr{I}^{-1}(\theta)/T, \\ \mathscr{I}(\theta) &= -\mathrm{E}\frac{\partial^2 L}{\partial\theta\,\partial\theta'}(\theta)/T. \end{aligned} \tag{7}$$

If $\hat{\theta}$ has a limiting normal distribution, and if $\mathscr{I}(\theta)$ is consistently estimated by $\mathscr{I}(\hat{\theta})$, then

$$\xi_W = T(\hat{\theta} - \theta^0)'\mathscr{I}(\hat{\theta})(\hat{\theta} - \theta^0) \tag{8}$$

will have a limiting X^2 distribution with k degrees of freedom when the null hypothesis is true. This is the Wald test based upon Wald's elegant (1943) analysis of the general asymptotic testing problem. It is the asymptotic approximation to the very familiar t and F tests in econometrics.

The likelihood ratio test is based upon the difference between the maximum of the likelihood under the null and under the alternative hypotheses. Under general conditions, the statistic,

$$\xi_{\mathrm{LR}} = -2\big(L(\theta^0, y) - L(\hat{\theta}, y)\big), \tag{9}$$

can be shown to have a limiting X^2 distribution under the null. Perhaps Wilks (1938) was the first to derive this general limiting distribution.

The Lagrange Multiplier test is derived from a constrained maximization principle. Maximizing the log-likelihood subject to the constraint that $\theta = \theta^0$ yields a set of Lagrange Multipliers which measure the shadow price of the constraint. If the price is high, the constraint should be rejected as inconsistent with the data. Letting H be the Lagrangian:

$$H = L(\theta, y) - \lambda'(\theta - \theta^0),$$

the first-order conditions are:

$$\frac{\partial L}{\partial \theta} = \lambda; \qquad \theta = \theta^0,$$

so $\lambda = s(\theta^0, y)$. Thus the test based upon the Lagrange Multipliers by Aitcheson and Silvey (1958) and Silvey (1959) is identical to that based upon the score as originally proposed by Rao (1948). In each case the distribution of the score is easily found under the null since it will have mean zero and variance $\mathscr{I}(\theta^0)T$. Assuming a central limit theorem applies to the scores:

$$\xi_{\mathrm{LM}} = s'(\theta^0, y)'\mathscr{I}^{-1}(\theta^0)s(\theta^0, y)/T, \tag{10}$$

will again have a limiting X^2 distribution with k degrees of freedom under the null.

The three principles are based on different statistics which measure the distance between H_0 and H_1. The Wald test is formulated in terms of $\theta^0 - \hat{\theta}$, the LR test in terms of $L(\theta^0) - L(\hat{\theta})$, and the LM test in terms of $s(\theta^0)$. A geometric interpretation of these differences is useful.

With $k = 1$, Figure 3.1 plots the log-likelihood function against θ for a particular realization y.

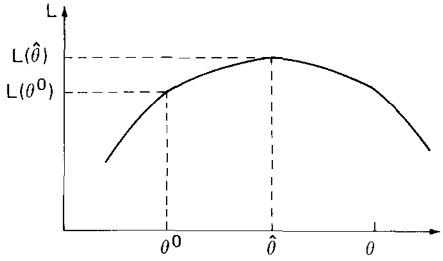

Figure 3.1

The MLE under the alternative is $\hat{\theta}$ and the hypothesized value is θ^0. The Wald test is based upon the horizontal difference between θ^0 and $\hat{\theta}$, the LR test is based upon the vertical difference, and the LM test is based on the slope of the likelihood function at θ^0. Each is a reasonable measure of the distance between H_0 and H_1 and it is not surprising that when L is a smooth curve well approximated by a quadratic, they all give the same test. This is established in Lemma 1.

Lemma 1

If $L = b - 1/2(\theta - \hat{\theta})'A(\theta - \hat{\theta})$ where A is a symmetric positive definite matrix which may depend upon the data and upon known parameters, b is a scalar and $\hat{\theta}$ is a function of the data, then the W, LR and LM tests are identical.

Proof

$$\partial L/\partial\theta = -(\theta - \hat{\theta})'A = s(\theta),$$
$$\partial^2 L/\partial\theta\,\partial\theta' = -A = -T\mathcal{I}.$$

Thus:

$$\xi_{\text{W}} = (\theta^0 - \hat{\theta})'A(\theta^0 - \hat{\theta}),$$
$$\begin{aligned}\xi_{\text{LM}} &= s(\theta^0)'A^{-1}s(\theta^0)\\ &= (\theta^0 - \hat{\theta})'A(\theta^0 - \hat{\theta}).\end{aligned}$$

Finally, by direct substitution:

$$\xi_{\text{LR}} = (\theta^0 - \hat{\theta})'A(\theta^0 - \hat{\theta}). \quad \text{Q.E.D.}$$

Whenever the true value of θ is equal or close to θ^0, then the likelihood function in the neighborhood of θ^0 will be approximately quadratic for large samples, with A depending only on θ^0. This is the source of the asymptotic equivalence of the tests for local alternatives and under the null which will be discussed in more detail in Section 6.

In the more common case where the null hypothesis is composite so that only a subset of the parameters are fixed under the null, similar formulae for the test statistics are available. Let $\theta = (\theta_1', \theta_2')'$ and $\hat{\theta} = (\hat{\theta}_1', \hat{\theta}_2')'$ where θ_1 is a $k_1 \times 1$ vector of parameters specified under the null hypothesis to be θ_1^0. The remaining parameters θ_2 are unrestricted under both the null and the alternative. The maximum likelihood estimate of θ_2 under the null is denoted $\tilde{\theta}_2$ and $\tilde{\theta} = (\theta_1^{0\prime}, \tilde{\theta}_2')'$.

Denote by $\mathscr{I}^{11}$ the partitioned inverse of $\mathscr{I}$ so that:

$$\mathscr{I}^{11^{-1}}=\mathscr{I}_{11}-\mathscr{I}_{12}\mathscr{I}_{22}^{-1}\mathscr{I}_{21}.$$

Then the Wald test is simply:

$$\xi_{\mathrm{W}}=T(\hat{\theta}_1-\theta_1^0)'\mathscr{I}^{11^{-1}}(\hat{\theta}_1-\theta_1^0), \tag{11}$$

which has a limiting X^2 distribution with k_1 degrees of freedom when H_0 is true. The LR statistic,

$$\xi_{\mathrm{LR}}=-2(L(\tilde{\theta},y)-L(\hat{\theta},y)), \tag{12}$$

has the same limiting distribution. The LM test is again derived from the Lagrangian:

$$H=L(\theta,y)-\lambda'(\theta_1-\theta_1^0),$$

which has first-order conditions:

$$\frac{\partial L}{\partial\theta_1}(\theta,y)=\lambda,$$
$$\frac{\partial L}{\partial\theta_2}(\theta,y)=0.$$

Thus:

$$\theta_1=\theta_1^0,$$
$$\xi_{\mathrm{LM}}=s(\tilde{\theta},y)'\mathscr{I}^{-1}(\tilde{\theta})s(\tilde{\theta},y)/T=s_1(\tilde{\theta},y)'\mathscr{I}^{11}s_1(\tilde{\theta},y)/T, \tag{13}$$

is the LM statistic which will again have a limiting X^2 distribution with k_1 degrees of freedom under the null. In Lemma 2 it is shown that again for the quadratic likelihood function, all three tests are identical.

Lemma 2

If the likelihood function is given as in Lemma 1 then the tests in (11), (12), and (13) are identical.

Proof

$$\begin{aligned}\xi_{\mathrm{W}}&=(\theta_1^0-\hat{\theta}_1)'A^{11^{-1}}(\theta_1^0-\hat{\theta}_1)\\&=(\theta_1^0-\hat{\theta}_1)'(A_{11}-A_{12}A_{22}^{-1}A_{21})(\theta_1^0-\hat{\theta}_1).\end{aligned}$$

For the other two tests, $\tilde{\theta}_2$ must be estimated. This is done simply by setting $S_2(\theta, y) = 0$:

$$\begin{pmatrix} S_1 \\ S_2 \end{pmatrix} = \frac{\partial L}{\partial \theta} = A(\theta - \hat{\theta}) = \begin{bmatrix} A_{11}(\theta_1 - \hat{\theta}_1) + A_{12}(\theta_2 - \hat{\theta}_2) \\ A_{21}(\theta_1 - \hat{\theta}_1) + A_{22}(\theta_2 - \hat{\theta}_2) \end{bmatrix} = 0.$$

So, $S_2 = 0$ implies:

$$\tilde{\theta}_2 - \hat{\theta}_2 = -A_{22}^{-1}A_{21}(\theta_1 - \hat{\theta}_1).$$

The concentrated likelihood function becomes:

$$L = b - \tfrac{1}{2}(\theta_1 - \hat{\theta}_1)'(A_{11} - A_{12}A_{22}^{-1}A_{21})(\theta_1 - \hat{\theta}_1),$$

and hence

$$\xi_{\mathrm{LR}} = (\theta_1^0 - \hat{\theta}_1)(A_{11} - A_{12}A_{22}^{-1}A_{21})(\theta_1^0 - \hat{\theta}_1).$$

Finally, the score is given by:

$$\begin{aligned} S_1(\tilde{\theta}) &= A_{11}(\theta_1^0 - \hat{\theta}_1) + A_{12}(\tilde{\theta}_2 - \hat{\theta}_2) \\ &= (A_{11} - A_{12}A_{22}^{-1}A_{21})(\theta_1^0 - \hat{\theta}_1). \end{aligned}$$

So

$$\xi_{\mathrm{LM}} = (\theta_1^0 - \hat{\theta}_1)'(A_{11} - A_{12}A_{22}^{-1}A_{21})(\theta_1^0 - \hat{\theta}_1). \quad \text{Q.E.D.}$$

Examination of the tests in (11), (12), and (13) indicates that neither the test statistic nor its limiting distribution under the null depends upon the value of the nuisance parameters θ_2. Thus the tests are (asymptotically) similar. It is apparent from the form of the tests as well as the proof of the lemma, that an alternative way to derive the tests is to first concentrate the likelihood function with respect to θ_2 and then apply the test for a simple null directly. This approach makes clear that by construction the tests will not depend upon the true value of the nuisance parameters. If the parameter vector has a joint normal limiting distribution, then the marginal distribution with respect to the parameters of interest will also be normal and the critical region will not depend upon the nuisance parameters either. Under general conditions therefore, the Wald, Likelihood Ratio and Lagrange Multiplier tests will be (asymptotically) similar.

As was described above, each of the tests can be thought of as depending on a statistic which measures deviations between the null and alternative hypotheses, and its distribution when the null is true. For example, the LM test is based upon the score whose limiting distribution is generally normal with variance $(\theta^0)\cdot T$ under the null. However, it is frequently easier to obtain the limiting distribution of the score in some other fashion and base the test on this. If a matrix V can be found so that:

$$T^{-1/2}s(\theta^0, y) \xrightarrow{D} N(0, V)$$

under H_0, then the test is simply:

$$\xi_{\mathrm{LM}} = s'V^{-1}s/T.$$

Under certain non-standard situations V may not equal $\mathscr{I}$ but in general it will. This is the approach taken by Engle (1982) which gives some test statistics very easily in complex problems.

4. Two simple examples

In these two examples, exact tests are available for comparison with the asymptotic tests under consideration.

Consider a set of T independent observations on a Bernoulli random variable which takes on the values:

$$y_t = \begin{cases} 1, & \text{with probability } \theta, \\ 0, & \text{with probability } 1-\theta. \end{cases} \tag{14}$$

The investigator wishes to test $\theta = \theta^0$ against $\theta \neq \theta^0$ for $\theta \in (0,1)$. The mean $\bar{y} = \sum y_t/T$ is a sufficient statistic for this problem and will figure prominently in the solution.

The log-likelihood function is given by:

$$L(\theta, y) = \sum_t (y_t \log\theta + (1-y_t)\log(1-\theta)), \tag{15}$$

with the maximum likelihood estimator, $\hat{\theta} = \bar{y}$. The score is:

$$s(\theta, y) = \frac{1}{\theta(1-\theta)} \sum_t (y_t - \theta).$$

Notice that $y_t - \theta$ is analogous to the "residual" of the fit. The information is:

$$\mathscr{I}(\theta) = \mathrm{E}\left[\frac{T\theta(1-\theta)+(1-2\theta)\Sigma(y_t-\theta)}{\theta^2(1-\theta)^2}\right]\Big/ T$$
$$= \frac{1}{\theta(1-\theta)}.$$

The Wald test is given by:

$$\xi_{\mathrm{W}} = T(\theta^0 - \bar{y})^2/\bar{y}(1-\bar{y}). \tag{16}$$

The LM test is:

$$\xi_{\mathrm{LM}} = \left[\frac{\Sigma(y_t-\theta^0)}{\theta^0(1-\theta^0)}\right]^2 \frac{\theta^0(1-\theta^0)}{T},$$

which is simply:

$$\xi_{\mathrm{LM}} = T(\theta^0 - \bar{y})^2/\theta^0(1-\theta^0). \tag{17}$$

Both clearly have a limiting chi-square distribution with one degree of freedom. They differ in that the LM test uses an estimate of the variance under the null whereas the Wald uses an estimate under the alternative. When the null is true (or a local alternative) these will have the same probability limit and thus for large samples the tests will be equivalent. If the alternative is not close to the null, then presumably both tests would reject with very high probability for large samples; the asymptotic behavior of tests for non-local alternatives is usually not of particular interest.

The likelihood ratio test statistic is given by:

$$\xi_{\mathrm{LR}} = 2T\{\bar{y}\log\bar{y}/\theta^0 + (1-\bar{y})\log(1-\bar{y})/(1-\theta^0)\}, \tag{18}$$

which has a less obvious limiting distribution and is slightly more awkward to calculate. A two-term Taylor series expansion of the statistic about $\bar{y} = \theta^0$ establishes that under the null the three will have the same distribution.

In each case, the test statistic is based upon the sufficient statistic $\bar{y}$. In fact, in each case the test is a monotonic function of $\bar{y}$ and therefore, the limiting chi squared approximation is not necessary. For each test statistic, the exact critical values can be calculated. Consequently, when the sizes of the tests are equal their critical regions will be identical; they will each reject for large values of $(\bar{y} - \theta^0)^2$.

The notion of how large it should be will be determined from the exact Binomial tables.

The second example is more useful to economists but has a similar result. In the classical linear regression problem, the test statistics are different, however, when corrected to have the same size they are identical for finite samples as well as asymptotically.

Let y^* and x^* be $T\times 1$ and $T\times k$ matrices satisfying:

$$y^*|x^* \sim N(x^*\beta, \sigma^2 I), \tag{19}$$

and consider testing the hypothesis that $R\beta = r$ where R is a $k_1\times k$ matrix of known constants and r is a $k_1\times 1$ vector of constants. If R has rank k_1, then the parameters and the data can always be rearranged so that the test is of omitted variable form. That is, (19) can be reparameterized in the notation of (4) as:

$$y|x \sim N(x\theta, \sigma^2 I), \tag{20}$$

where the null hypothesis is $\theta_1 = 0$ and y and x are linear combinations of y^* and x^*. In this particular problem it is just as easy to use (19) as (20); however, in others the latter form will be simpler. The intuitions are easier when the parameters of R and r do not appear explicitly in the test statistics. Furthermore, (20) is most often the way the test is calculated to take advantage of packaged computer programs since it involves running regressions with and without the variables x_1.

For the model in (20) the log-likelihood conditional on x is:

$$L(\theta, y) = k - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - x\theta)'(y - x\theta), \tag{21}$$

where k is a constant. If σ^2 were known, Lemmas 1 and 2 would guarantee that the W, LR, and LM tests would be identical. Hence, the important difference between the test statistics will be the estimate of σ^2. The score and information matrix corresponding to the parameters θ are:

$$\begin{aligned} s(\theta, y) &= x'u/\sigma^2; \qquad u = y - x\theta, \\ \mathcal{I}_{\theta\theta} &= x'x/\sigma^2 T, \end{aligned} \tag{22}$$

and the information matrix is block diagonal between θ and σ^2. Notice that the score is proportional to the correlation coefficient between the residuals and the x variables. This is of course zero at $\hat{\theta}$ but not at the estimates under the null, $\tilde{\theta}$.

The three test statistics therefore are:

$$\xi_W = \left(\theta_1^0 - \hat{\theta}_1\right)'\left(x_1'x_1 - x_1'x_2\left(x_2'x_2\right)^{-1}x_2'x_1\right)\left(\theta_1^0 - \hat{\theta}_1\right)/\hat{\sigma}^2, \tag{23}$$

$$\xi_{LM} = \tilde{u}'x_1\left(x_1'x_1 - x_1'x_2\left(x_2'x_2\right)^{-1}x_2'x_1\right)^{-1}x_1'\tilde{u}/\tilde{\sigma}^2, \tag{24}$$

$$\xi_{LR} = T\log(\tilde{u}'\tilde{u}/\hat{u}'\hat{u}), \tag{25}$$

where $\hat{u} = y - x\hat{\theta}$, $\tilde{u} = y - x\tilde{\theta}$, and $\hat{\sigma}^2 = \hat{u}'\hat{u}/T$, $\tilde{\sigma}^2 = \tilde{u}'\tilde{u}/T$, and x is conformably partitioned as $x = (x_1, x_2)$. From the linear algebra of projections, these can be rewritten as:

$$\xi_W = T(\tilde{u}'\tilde{u} - \hat{u}'\hat{u})/\hat{u}'\hat{u}, \tag{26}$$

$$\xi_{LM} = T(\tilde{u}'\tilde{u} - \hat{u}'\hat{u})/\tilde{u}'\tilde{u}. \tag{27}$$

This implies that:

$$\xi_{LR} = T\log(1 + \xi_W/T); \qquad \xi_{LM} = \xi_W/(1 + \xi_W/T),$$

and that $(T - K)\xi_W/TK_1$ will have an exact $F_{k_1, T-k}$ distribution under the null. As all the test statistics are monotonic functions of the F statistic, then exact tests for each would produce identical critical regions. If, however, the asymptotic distribution is used to determine the critical values, then the tests will differ for finite samples and there may be conflicts between their conclusions. Evans and Savin (1980) calculate the probabilities of such conflicts for the test in (23)–(25) as well as for those modified either by a degree of freedom correction or by an Edgeworth expansion correction. In the latter case, the sizes are nearly correct and the probability of conflict is nearly zero. It is not clear how these conclusions generalize to models for which there are no exact results but similar conclusions might be expected. See Rothenberg (1980) for some evidence for the equivalence of the tests for Edgeworth expansions to powers of $1/T$.

5. The linear hypothesis in generalized least squares models

5.1. *The problem*

In the two preceding examples, there was no reason to appeal to asymptotic approximations for test statistics. However, if the assumptions are relaxed slightly, then the exact tests are no longer available. For example, if the variables were simply assumed contemporaneously uncorrelated with the disturbances as in:

$$y_t|x_t \sim \mathrm{IN}(x_t\beta, \sigma^2), \tag{28}$$

where IN means independent normal, then the likelihood would be identical but the test statistics would not be proportional to an F distributed random variable. Thus, inclusion of lagged dependent variables or other predetermined variables would bring asymptotic criteria to the forefront in choosing a test statistic and any of the three would be reasonable candidates as would the standard F approximations. Similarly, if the distribution of y is not known to be normal, a central limit theorem will be required to find the distribution of the test statistics and therefore only asymptotic tests will be available.

The important case to be discussed in this section is testing a linear hypothesis when the model is a generalized least squares model with unknown parameters in the covariance matrix. Suppose:

$$y|x \sim N(x\beta, \sigma^2\Omega), \qquad \Omega = \Omega(\omega), \tag{29}$$

where ω is a finite estimable parameter vector. The model has been formulated so that the hypothesis to be tested is H_0: $\beta_1 = 0$, where $\beta = (\beta_1', \beta_2')'$ and x is conformally partitioned as $x = (x_1, x_2)$. The collection of parameters is now $\theta = (\beta_1', \beta_2', \sigma^2, \omega')'$.

A large number of econometric problems fit into this framework. In simple linear regression the standard heteroscedasticity and serial correlation covariance matrices have this form. More generally if ARMA processes are assumed for the disturbances or they are fit with spectral methods assuming only a general stationary structure as in Engle (1980), the same analysis will apply. From pooled time series of cross sections, variance component structures often arise which have this form. To an extent which is discussed below, instrumental variables estimation can be described in this framework. Letting X be the matrix of all instruments, $X(X'X)^{-1}X'$ has no unknown parameters but acts like a singular covariance matrix. Because it is an idempotent matrix, its generalized inverse is just the matrix itself, and therefore many of the same results will apply.

For systems of equations, a similar structure is often available. By stacking the dependent variables in a single dependent vector and conformably stacking the independent variables and the coefficient vectors, the covariance matrix of a seemingly unrelated regression problem (SUR) will have a form satisfied by (29). In terms of tensor products this covariance matrix is $\Omega = \Sigma \otimes I$, where Σ is the contemporaneous covariance matrix. Of course more general structures are also appropriate. The three stage least squares estimator also is closely related to this analysis with a covariance matrix $\Omega = \Sigma \otimes X(X'X)^{-1}X'$.

5.2. *The test statistics*

The likelihood function implied by (29) is given by:

$$L(\theta, y) = k - \frac{T}{2}\log\sigma^2 - \tfrac{1}{2}\log|\Omega| - \frac{1}{2\sigma^2}(y - x\beta)'\Omega^{-1}(y - x\beta). \tag{30}$$

Under these assumptions it can be shown that the information matrix is block diagonal between the parameters β and (σ^2, ω). Therefore attention can be confined to the β components of the score and information. These are given by:

$$s_{\beta_1}(\theta, y) = x_1'\Omega^{-1}u/\sigma^2, \qquad u = y - x\beta, \tag{31}$$

$$\mathscr{I}_{\beta\beta}(\theta) = x'\Omega^{-1}x/\sigma^2 T. \tag{32}$$

Denote the maximum likelihood estimates of the parameters under H_1 by $\hat{\theta} = (\hat{\beta}, \hat{\sigma}^2, \hat{\omega})$ and let $\hat{\Omega} = \Omega(\hat{\omega})$; denote the maximum likelihood estimates of the same parameters under the null as $\tilde{\theta} = (\tilde{\beta}, \tilde{\sigma}^2, \tilde{\omega})$ and let $\tilde{\Omega} = \Omega(\tilde{\omega})$. Further, let $\hat{u} = y - x\hat{\beta}$ and $\tilde{u} = y - x\tilde{\beta}_2$ be residuals under the alternative and the null.

Then substituting into (11), (12), and (13), the test statistics are simply:

$$\xi_{\mathrm{W}} = \hat{\beta}_1'\left(x_1'\hat{\Omega}^{-1}x_1 - x_1'\hat{\Omega}^{-1}x_2\left(x_2'\hat{\Omega}^{-1}x_2\right)^{-1}x_2'\hat{\Omega}^{-1}x_1\right)\hat{\beta}_1/\hat{\sigma}^2, \tag{33}$$

$$\xi_{\mathrm{LR}} = -2(L(\tilde{\theta}, y) - L(\hat{\theta}, y)), \tag{34}$$

$$\xi_{\mathrm{LM}} = \tilde{u}'\tilde{\Omega}^{-1}x_1\left(x_1'\tilde{\Omega}^{-1}x_1 - x_1'\tilde{\Omega}^{-1}x_2\left(x_2'\tilde{\Omega}^{-1}x_2\right)^{-1}x_2'\tilde{\Omega}^{-1}x_1\right)^{-1}x_1'\tilde{\Omega}^{-1}\tilde{u}/\tilde{\sigma}^2. \tag{35}$$

The Wald statistic can be recognized as simply the F or squared t statistic commonly computed by a GLS regression (except for finite sample degree of freedom corrections). This illustrates that for testing one parameter, the square root of these statistics with the appropriate sign would be the best statistic since it would allow one tailed tests if these are desired.

It is well known that the Wald test statistic can be calculated by running two regressions just as in (26). Care must however be taken to use the same metric (estimate of Ω) for both the restricted and the unrestricted regressions. The residuals from the unrestricted regression using $\hat{\Omega}$ as the covariance matrix are the $\hat{u}$, however, the residuals from the restricted regression using $\hat{\Omega}$ are not $\hat{u}$. Let them be denoted u^{01} indicating the model under H^0 with the covariance matrix under H^1. Thus, $u^{01} = y - x_2\beta_2^{01}$ is calculated assuming $\hat{\Omega}$ is a known matrix. The Wald statistic can equivalently be written as:

$$\xi_{\mathrm{W}} = T(u^{01\prime}\hat{\Omega}^{-1}u^{01} - \hat{u}'\hat{\Omega}^{-1}\hat{u})/\hat{u}'\hat{\Omega}^{-1}\hat{u}. \tag{36}$$

The LM statistic can also be written in several different forms some of which may be particularly convenient. Three different versions will be given below.

Because $\tilde{u}'\tilde{\Omega}^{-1}x_2 = 0$ by the definition of $\tilde{u}$, the LM statistic is more simply written as:

$$\xi_{\mathrm{LM}} = T\tilde{u}'\tilde{\Omega}^{-1}x(x'\tilde{\Omega}^{-1}x)^{-1}x'\tilde{\Omega}^{-1}\tilde{u}/\tilde{u}'\tilde{\Omega}^{-1}\tilde{u}. \tag{37}$$

This can be interpreted as T times the R^2 of a regression where $\tilde{u}$ is the dependent variable, x is the set of independent variables and $\tilde{\Omega}$ is the covariance matrix of the disturbances which is assumed known. From the formula it is clear that this should be the R^2 calculated as the explained sum of squares over the total sum of squares. This is in contrast to the more conventional measure where these sums of squares are about the means. Furthermore, it is clear that the data should first be transformed by a matrix P such that $P'P = \tilde{\Omega}^{-1}$, and then the auxiliary regression and R^2 calculated. As there may be ambiguities in the definition of R^2 when $\Omega \neq I$ and when there is no intercept in the regression, let R_0^2 represent the figure implied by (37). Then:

$$\xi_{\mathrm{LM}} = TR_0^2. \tag{38}$$

In most cases and for most computer packages R_0^2 will be the conventionally measured R^2. In particular when Px includes an intercept under H_0, then $P\hat{u}$ will have a zero mean so that the centered and uncentered sums of squares will be equal. Thus, if the software first transforms the data by P, the R^2 will be R_0^2.

A second way to rewrite the LM statistic is available along the lines of (27). Let u^{10} be the residuals from a regression of y on the unrestricted model using $\tilde{\Omega}$ as the covariance matrix, so that $u^{10} = y - x\beta^{10}$. Then the LM statistic is simply:

$$\xi_{\mathrm{LM}} = T(\tilde{u}'\tilde{\Omega}^{-1}\tilde{u} - u^{10\prime}\tilde{\Omega}^{-1}u^{10})/\tilde{u}'\tilde{\Omega}^{-1}\tilde{u}. \tag{39}$$

A statistic which differs only slightly from the LM statistic comes naturally out of the auxiliary regression. The squared t or F statistics associated with the variables x_1 in the auxillary regressions of $\tilde{u}$ on x using $\tilde{\Omega}$ are of interest. Letting:

$$A = x_1'\tilde{\Omega}^{-1}x_1 - x_1'\tilde{\Omega}^{-1}x_2(x_2'\tilde{\Omega}^{-1}x_2)^{-1}x_2'\tilde{\Omega}^{-1}x_1,$$

then

$$\beta^{10} = (x'\tilde{\Omega}^{-1}x)^{-1}x'\tilde{\Omega}^{-1}\tilde{u},$$

or the first elements $\beta_1^{10} = A^{-1}x_1'\tilde{\Omega}^{-1}\tilde{u}$. The F statistic aside from degree of freedom corrections is given by:

$$\begin{aligned}\xi'_{\mathrm{LM}} &= \beta_1^{10\prime}A\beta_1^{10}/\sigma^{2(10)}\\ &= \tilde{u}'\tilde{\Omega}^{-1}x_1A^{-1}x_1'\tilde{\Omega}^{-1}\tilde{u}/\sigma^{2(10)},\end{aligned} \tag{40}$$

where $\sigma^{2(10)}$ is the residual variance from this estimation. From (35) it is clear that $\xi_{\mathrm{LM}} = \xi'_{\mathrm{LM}}$ if $\sigma^{2(10)} = \tilde{\sigma}^2$. The tests will differ when x_1 explains some of $\tilde{u}$, that is, when H_0 is not true. Hence, under the null and local alternatives, these two variances will have the same probability limit and therefore the tests will have the same limiting distribution. Furthermore, adding a linear combination of regressors to both sides of a regression will not change the coefficients or the significance of other regressors. In particular adding $x_2\tilde{\beta}_2$ to both sides of the auxiliary regression converts the dependent variable to y and yet will not change ξ'_{LM}. Hence, the t or F tests obtained from regressing y on x_1 and x_2 using $\tilde{\Omega}$ will be asymptotically equivalent to the LM test.

5.3. *The inequality*

The relationship between the Wald and LM tests in this context is now clearly visible in terms of the choice of Ω to use for the test. The Wald test uses $\hat{\Omega}$ while the LM test uses $\tilde{\Omega}$ and the Likelihood Ratio test uses both. As the properties of the tests differ only for finite samples, frequently computational considerations will determine which to use. The primary computational differences stem from the estimation of Ω which may require non-linear or other iterative procedures. It may further require some specification search over a class of possible disturbance specifications. The issue therefore hinges upon whether $\hat{\Omega}$ or $\tilde{\Omega}$ is already available from previous calculations. If the null hypothesis has already been estimated and the investigator is trying to determine whether an additional variable belongs in the model in the spirit of diagnostic testing, then $\tilde{\Omega}$ is already estimated and the LM test is easier. If on the other hand, the more general model has been estimated, and the test is for a simplification or a test of a theory which predicts the importance of some variable, then $\hat{\Omega}$ is available and the Wald test is easier. In rare cases will the LR test be computationally easier.

The three test statistics differ for finite samples but are asymptotically equivalent. When the critical regions are calculated from the limiting distributions, then there may be conflicts in inference between the tests. The surprising character of this conflict is pointed out by a numerical inequality among the test statistics. It was originally established by Savin (1976) and Berndt and Savin (1977) for special cases of (29) and then by Breusch (1979) in the general case of (29). For any data set y, x, the three test statistics will satisfy the following inequality:

$$\xi_{\mathrm{W}} \geq \xi_{\mathrm{LR}} \geq \xi_{\mathrm{LM}}. \tag{41}$$

Therefore, whenever the LM test rejects, so will the others and whenever the W fails to reject, so do the others. The inequality, however, has nothing to say about the relative merits of the tests because it applies under the null as well. That is, if the Wald test has a size of 5%, then the LR and LM test will have a size less than 5%. Hence their apparently inferior power performance is simply a result of a more conservative size. When the sizes are corrected to be the same, there is no longer a simple inequality relationship on the powers. As mentioned earlier, both Rothenberg (1979) and Evans and Savin (1982) present results that when the sizes are approximately corrected, the powers are approximately the same.

5.4. *A numerical example*

As an example, consider an equation presented in Engle (1978) which explains employment in Boston's textile industry as a function of the U.S. demand and prices, the stock of fixed factors in Boston and the Boston wage rate. The equation is a reduced form derived from a simple production model with capital as a fixed factor and a constant price elasticity of demand. The variables are specific combinations of logarithms of the original data. Denote the dependent variable by y, and the independent variables by x_1, x_2 and a constant. The hypothesis to be tested is whether a time trend should also be introduced to allow technical progress in the sector. There is substantial serial correlation in the disturbance and several methods of parameterizing it are given in the original paper; however, it will here be assumed to follow a first-order autoregressive process. There are 22 annual observations.

The basic estimate of the relation is:

$$\tilde{y} = \underset{(0.92)}{4.4} + \underset{(2.45)}{0.165x_1} + \underset{(3.11)}{0.669x_2}; \qquad \rho = 0.901, \quad R^2 = 0.339.$$

The estimate is not particularly good but it has the right signs and significant t-statistics. Rho was estimated by searching over the unit interval and the estimate is maximum likelihood.

The residuals from this estimate were then regressed upon the expanded set of regressors, to obtain:

$$\tilde{u} = \underset{(1.90)}{49.2} - \underset{(-1.61)}{0.185x_1} - \underset{(-0.22)}{0.045x_2} - \underset{(1.93)}{0.025\text{ time}}; \qquad \rho = 0.901, \quad R^2 = 0.171.$$

The same value of rho was imposed upon this estimate. The Lagrange Multiplier statistic is (22) (0.171) = 3.76 which is slightly below the 95% level for $X_1^2(3.84)$ but above the 90% level (2.71) so it rejects at 90% but not 95%. Notice that the t-statistic on time is not significant at 95% but is at the 90% level.

For comparison, the full regression was estimated including a reoptimization of rho. The results were

$$\hat{y} = \underset{(2.26)}{59.9} - \underset{(-0.45)}{0.05x_1} + \underset{(3.18)}{0.611x_2} - \underset{(2.13)}{0.028} \text{ time}; \qquad \rho = 0.970, \quad R^2 = 0.480.$$

The Wald test involves merely looking at the t-statistic on time; however, the asymptotic formulation would estimate the standard error using 22 degrees of freedom rather than 18. In this case the t-statistic is -2.35 so the test rejects at 95% but not 99%. The Wald statistic $\xi_W = 5.52$ exceeds the 95 point of X_1^2 but not the 99% point (6.63).

In this example the two test statistics give conflicting inference at the 95% level with the Wald statistic rejecting the null hypothesis and the Lagrange Multiplier statistic accepting. However, at both 90% and the 99% level, they agree. The numerical results support the algebraic relationship given above. The benefits from using the Lagrange Multiplier test lie primarily in the avoidance of a recalculation of rho. While this may appear a rather minimal saving for the first-order autoregressive case, it may be substantial for models postulated to have ARMA disturbance processes or general stationary error processes requiring expensive iterative procedures. In establishing the validity of a regression equation, a variety of alternatives may be considered and thus, the computational saving from such a battery of tests will be even more substantial.

5.5. *Instrumental variables*

A closely related set of problems occurs in testing hypotheses in equations or models estimated with instrumental variables methods. The analysis given here concerns only the Wald test in several forms, however, LM versions can be deduced from the results in Engle (1982).

Consider first a single equation in a simultaneous system:

$$y = Y\alpha + x\gamma + \varepsilon = z\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I), \tag{42}$$

and X is a matrix of instrumental variables including x, which is assumed to be uncorrelated with ε but correlated with Y. Limited information maximum likelihood estimation of this model yields asymptotically the same estimates as 2SLS or IV, and hence the standard test statistics are asymptotically equivalent to Wald tests. Letting $G = X(X'X)^{-1}X'$ and H_0; $\beta_1 = 0$ be the hypothesis under test, the standard test statistic is simply:

$$\xi'_W = \hat{\beta}_1'\left(z_1'Gz_1 - z_1'Gz_2\left(z_2'Gz_2\right)^{-1}z_2'Gz_1\right)\hat{\beta}_1/\hat{\sigma}^2, \tag{43}$$

where $\hat{\beta}=(z'Gz)^{-1}z'Gy$, $\hat{u}=y-z\hat{\beta}$, $\hat{\sigma}^2=\hat{u}'\hat{u}/T$. This expression is identical to that in (36) except that the estimates of σ^2 are different. In (36) $\hat{\sigma}^2=\hat{u}'\Omega^{-1}\hat{u}/T$ instead of $\hat{u}'\hat{u}/T$. Following the line of reasoning leading to (37), the numerator can be rewritten in terms of the residuals from a restricted regression using the same G matrix. Letting $\tilde{\beta}_2=(z_2'Gz_2)^{-1}z_2'Gy$ and $\tilde{u}=y-z_2\tilde{\beta}_2$, the statistic can be expressed as:

$$\xi_{\mathrm{W}}'=T(\tilde{u}'G\tilde{u}-\hat{u}'G\hat{u})/\hat{u}'\hat{u}. \tag{44}$$

Because G is idempotent, the two sums of squares in the numerator can be calculated by regressing the corresponding residuals on X and looking at the explained sums of squares. Their difference is also available as the difference between the sums of squared residuals from the second stages of the relevant 2SLS regressions.

As long as the instrument list is unchanged from the null to the alternative hypothesis, there is no difficulty formulating this test. If the list does change then the Wald test appropriately uses the list under the alternative. One might suspect that a similar LM test would be available using the more limited set of instruments, however, this is not the case at least in this simple form. When the instruments are different, the LM test can be computed as given in Engle (1979a) but does not have the desired simple form.

In the more general case where (42) represents a stacked set of simultaneous equations the covariance would in general be given by $\Sigma\otimes I$, where Σ is the contemporaneous covariance matrix. The instruments in the stacked system can be formulated as $I\otimes X$ and therefore letting $\hat{\Sigma}$ be the estimated covariance matrix under the alternative, the 3SLS estimator can be written letting $G=\hat{\Sigma}\otimes X(X'X)^{-1}X'$ as:

$$\hat{\beta}=(z'Gz)^{-1}z'Gy.$$

Again, through the equivalence with FIML, the approximate Wald test is:

$$\xi_{\mathrm{W}}'=\hat{\beta}_1'\left(z_1'Gz_1-z_1'Gz_2\left(z_2'Gz_2\right)^{-1}z_2'Gz_1\right)\hat{\beta}_1,$$

which can be reformulated as:

$$=T(\tilde{u}'G\tilde{u}-\hat{u}'G\hat{u}).$$

Notice that $\hat{\sigma}^2$ has disappeared from the test statistic as it is incorporated in G through $\hat{\Sigma}$. Again this difference is equal to the difference between the sums of squared residuals in the restricted and unrestricted third stage of 3SLS.

6. Asymptotic equivalence and optimality of the test statistics

In this section the asymptotic equivalence, the limiting distributions and the asymptotic optimality of the three test statistic will be established under the conditions of Crowder (1976). These rather weak conditions allow some dependence of the observations and do not require that they be identically distributed. Most econometric problems will be encompassed under these assumptions. Although it is widely believed that these tests are optimal in some sense, the discussion in this section is designed to establish their properties under a set of regularity conditions.

The log likelihood function assumed by Crowder allows for general dependence of the random variables and for some types of stochastic or deterministic exogenous variables. Let $Y_0, Y_1, \dots, Y_T$ be $p \times 1$ vectors of random variables which have known conditional probability density functions $f_t(Y_t|\mathscr{F}_{t-1}; \theta)$, where $\theta \in \Theta$ an open subset of R^k and $\mathscr{F}_{t-1}$ is the σ field generated by $Y_0, \dots, Y_{t-1}$, the "previous history". The log-likelihood conditional on Y_0 is:

$$L_T(Y;\theta) = \sum_{t=1}^{T} \log f_t(Y_t|\mathscr{F}_{t-1}, \theta). \tag{45}$$

In this expression, non-stochastic variables enter through the time subscript on f which allows each random vector to be distributed differently. Stochastic variables which appear in conditioning sets can also be included within this framework if they satisfy the assumptions of weak exogeneity as defined by Engle, Hendry and Richard (1983). Let $Y_t = (y_t, x_t)$, where the parameters of the conditional distribution of y given x, $g_t(y_t|x_t, \mathscr{F}_{t-1}, \theta)$ are of interest. Then expressing the density of x as $h_t(x_t|\mathscr{F}_{t-1}, \phi)$ for some parameters ϕ, the log-likelihood function can be written as:

$$L_T(Y,\theta,\phi) = \sum_{t=1}^{T} \log g_t(y_t|x_t, \mathscr{F}_{t-1}, \theta) + \sum_{t=1}^{T} \log h_t(x_t|\mathscr{F}_{t-1}, \phi).$$

If ϕ is irrelevant to the analysis, then x_t is weakly exogenous. The information matrix will clearly be block diagonal between θ and ϕ and the MLE of θ will be obtained just by maximizing the first sum with respect to θ. Therefore, if the log-likelihood L_T satisfies Crowder's assumptions, then the conditional log-likelihood,

$$L_T^*(y,x,\theta) = \sum_{t=1}^{T} \log g_t(y_t|x_t, \mathscr{F}_{t-1}, \theta),$$

also will. Notice that this result requires only that x be weakly exogenous; it need not be strongly exogenous and can therefore depend upon past values of y.

The GLS models of Section 5 can now also be written in this framework. Letting $P'P = \Omega^{-1}$ for any value of ω, rewrite the model with $y^* = Py$, $x^* = Px$ so that:

$$y^*|x^* \sim N(x^*\beta, \sigma^2 I)$$

The parameters of interest are now β, σ^2 and ω. If the x were fixed constants, then so will be the x^*. If the x were stochastic strongly exogenous variables as implied by (29), then so will be x^*. The density $h(x, \phi)$ will become $h^*(x^*, \phi, \omega)$ but unless there is some strong a priori structure on h, ω will not enter h^*. If the covariance structure is due to serial correlation then rewriting the model conditional on the past will transform it directly into the Crowder framework regardless of whether the model is already dynamic or not.

Based on (45), the score, Hessian and information matrix are defined by:

$$\begin{aligned} s_T(y,\theta) &= \frac{\partial L(y,\theta)}{\partial\theta}, \\ H_T(y,\theta) &= \frac{\partial^2 L}{\partial\theta\,\partial\theta'}(y,\theta), \\ \mathscr{I}_T(\theta) &= \frac{1}{T}Es_T(y,\theta)s_T'(y,\theta). \end{aligned} \tag{46}$$

Notice that the information matrix depends upon the sample size because the y_t's are not identically distributed.

The essential conditions assumed by Crowder are:

(a) the true θ, θ^*, is an interior point of $\textcircled{\Theta}$;
(b) the Hessian matrix is a continuous function of θ in a neighborhood of θ^*;
(c) $\mathscr{I}_T(\theta^*)$ is non-singular;
(d) plim $(\mathscr{I}_T^{-1}(\theta)H_T(y,\theta)/T) = I$ for θ in a neighborhood of θ^*; and
(e) a condition such that no term in y_t dominates the sum to T.

Suppose the hypothesis to be tested is H_0: $\theta = \theta^0$ while the alternative is H_1: $\theta = \theta^T$ where plim $T^{1/2}(\theta^T - \theta^0) = \delta$ for some vector δ.

Under these assumptions the maximum likelihood estimator of $\theta, \hat{\theta}$ exists and is consistent with a limiting normal density given by:

$$T^{1/2}\mathscr{I}_T^{1/2}(\theta^*)(\hat{\theta} - \theta^*) \xrightarrow{D} N(0, I) \tag{47}$$

Mean Value Taylor series expansions can be written as:

$$L(\theta, y) = L(\hat{\theta}, y) - \frac{T}{2}(\theta - \hat{\theta})' A_T(\theta, \hat{\theta})(\theta - \hat{\theta}),$$
$$s_T(\theta, y) = -TA_T(\theta, \hat{\theta})(\theta - \hat{\theta}), \tag{48}$$

where $T[A_T(\theta, \hat{\theta})]_{ij} = [H_T(\bar{\theta})]_{ij}$ and $\bar{\theta} \in (\theta, \hat{\theta})$ possibly at different points for different (i, j). From (48) the Likelihood Ratio test is simply:

$$\xi_{\text{LR}} = T(\theta^0 - \hat{\theta})' A_T(\theta^0, \hat{\theta})(\theta^0 - \hat{\theta}),$$

and the Wald test is:

$$\xi_{\text{W}} = T(\theta^0 - \hat{\theta})' \mathscr{I}_T(\hat{\theta})(\theta^0 - \hat{\theta}).$$

Thus,

$$\text{plim}|\xi_{\text{LR}} - \xi_{\text{W}}| = \text{plim}|T(\theta^0 - \hat{\theta})'(A_T(\theta^0, \hat{\theta}) - \mathscr{I}_T(\hat{\theta}))(\theta^0 - \hat{\theta})|.$$

The plim of the middle terms is zero for $\theta^* = \theta^0$ and for the sequence of local alternatives since again plim $\theta^T = \theta^0$. The terms $T^{1/2}(\hat{\theta} - \theta^0)$ will converge in distribution under both H_0 and H_1 and therefore the product converges in probability to zero under H_0 and H_1. Thus ξ_{LR} and ξ_{W} have the same limiting distributions. Similarly, from (48) and (10):

$$\xi_{\text{LM}} = Ts_T(\theta^0, y)' \mathscr{I}_T(\theta^0)^{-1} s_T(\theta^0, y)$$
$$= T(\theta^0 - \hat{\theta})' A_T(\theta^0, \hat{\theta}) \mathscr{I}_T(\theta^0)^{-1} A_T(\theta^0, \hat{\theta})(\theta^0 - \hat{\theta}),$$

and by the same argument $\text{plim}|\xi_{\text{LR}} - \xi_{\text{LM}}| = 0$ for H_0 and local alternatives. Thus we have the following theorem:

Theorem 1

Under the assumptions in Crowder (1976), the Wald, Likelihood Ratio and Lagrange Multiplier test statistics have the same limiting distribution when the null hypothesis or local alternative are true.

Another way to describe this result is to rewrite (48) as:

$$L(\theta, y) = L(\hat{\theta}, y) - \frac{T}{2}(\theta - \hat{\theta})' \mathscr{I}_T(\theta^0)(\theta - \hat{\theta}) + O_p(1), \tag{49}$$

where $O_p(1)$ refers to the remainder terms which vanish in probability for H_0 and local alternatives. Thus, asymptotically the likelihood is exactly quadratic and Lemmas 1 and 2 establish that the tests are all the same. Furthermore, (49) establishes that $\hat{\theta}$ is asymptotically sufficient for θ. To see this more clearly, rewrite the joint density of y as:

$$f(y,\theta)=f(y,\hat{\theta})\exp\left[-\tfrac{1}{2}(\theta-\hat{\theta})'\mathscr{I}_T(\theta^0)(\theta-\hat{\theta})\right]+O_p(1)$$

and notice that by the factorization theorem, $\hat{\theta}$ is sufficient for θ as long as y does not enter the exponent which will be true asymptotically.

Finally, because $\hat{\theta}$ has a limiting normal distribution, with a known covariance matrix $\mathscr{I}(\theta^0)^{-1}$, all the testing results for hypotheses on the mean vector of a multivariate normal, now apply asymptotically by considering $\hat{\theta}$ as the data.

To explore the nature of this optimality, suppose that the likelihood function in (49) is exact without the $O_p(1)$ term. Then several results are immediately apparent. If θ is one dimensional, uniformly most powerful (UMP) tests will exist against one sided alternatives and UMP unbiased (UMPU) tests will exist against two sided alternatives.

If $\theta=(\theta_1,\theta_2)$ where θ_1 is a scalar hypothesized to have value θ_1^0 under H_0 but θ_2 are unrestricted, then UMP similar or UMPU tests are available.

When θ_1 is multivariate, an invariance criterion must be added. In testing the hypothesis $\mu=0$ in the canonical model $V\sim N(\mu, I)$, there is a natural invariance with respect to rotations of V. If $\tilde{V}=DV$, where D is an orthogonal matrix, then the testing problem is unchanged so that a test should be invariant to whether V or $\tilde{V}$ are given. Essentially, this invariance says that the test should not depend on which order the V's are in; it should be equally sensitive to deviations in all directions. The maximally invariant statistic in this problem is $\sum V_i^2$ which means that any test which is to be invariant can be based upon this statistic. Under the assumptions of the model, this will be distributed as $X_k^2(\lambda)$ with non-centrality parameter $\lambda=\mu'\mu$. The Neyman–Pearson lemma therefore establishes that the uniformly most powerful invariant test would be based upon a critical region:

$$C=\left\{\sum V_i^2>c\right\}.$$

To rewrite (49) in this form, let $\mathscr{I}_T(\theta^0)^{-1}=P'P$ and $V=P(\hat{\theta}-\theta^0)$. Then the maximal invariant is

$$T(\hat{\theta}-\theta^0)'\mathscr{I}_T(\theta^0)(\hat{\theta}-\theta^0)$$

which is distributed as $X_k^2(\lambda)$ where $\lambda=T\delta'\mathscr{I}_T(\theta^0)\delta$ where $\delta=\theta^1-\theta^0$. The non-centrality parameter depends upon the distance between the null and alternative hypotheses in the metric $\mathscr{I}_T(\theta^0)$.

If the null hypothesis in the canonical model specifies merely H_0: $\mu_1 = 0$, then an additional invariance argument is invoked, namely $\tilde{V}_2' = V_2 + K$, where K is an arbitrary set of constants, and $V' = (V_1', V_2')$. Then the maximal invariant is $V_1'V_1$ which in (49) becomes:

$$\xi = T(\hat{\theta}_1 - \theta_1^0)'(\mathscr{I}_{11} - \mathscr{I}_{12}\mathscr{I}_{22}^{-1}\mathscr{I}_{21})(\hat{\theta}_1 - \theta_1^0). \tag{50}$$

The non-centrality parameter becomes:

$$\lambda = \mu_1'\mu_1 = T\delta_1'(\mathscr{I}_{11} - \mathscr{I}_{12}\mathscr{I}_{22}^{-1}\mathscr{I}_{21})\delta_1. \tag{51}$$

Thus, any test which is invariant can be based on this statistic and a uniformly most powerful invariant test would have a critical region of the form:

$$C = \{\xi \geq c\}.$$

This argument applies directly to the Wald, Likelihood Ratio and LM tests. Asymptotically the remainder term in the likelihood function vanishes for the null hypothesis and for local alternatives. Hence, these tests can be characterized as asymptotically locally most powerful invariant tests. This is the general optimality property of such tests which often will be simply called asymptotic optimality. For further details on these arguments the reader is referred to Cox and Hinckley (1974, chs. 5, 9), Lehmann (1959, chs. 4, 6, 7), and Fergurson (1967, chs. 4, 5).

In finite samples many tests derived from these principles will have stronger properties. For example, if a UMP test exists, a locally most powerful test will be it. Because of the invariance properties of the likelihood function it will automatically generate tests with most invariance properties and all tests will be functions of sufficient statistics.

One further property of Lagrange Multiplier tests is useful as it gives a general optimality result for finite samples. For testing H_0: $\theta = \theta^0$ against a local alternative H_1: $\theta = \theta^0 + \delta$ for δ a vector of small numbers, the Neyman–Pearson lemma shows that the likelihood ratio is a sufficient statistic for the test. The likelihood ratio is:

$$\begin{aligned} e^{\lambda} &= L(\theta^0, y) - L(\theta^0 + \delta, y) \\ &= s(\theta^0, y)'\delta, \end{aligned}$$

for small δ. The best test for local alternatives is therefore based on a critical region:

$$C = \{ s'\delta > c \}.$$

In this case δ chooses a direction. However, if invariance is desired, then the test would be based upon the scores in all directions:

$$C = \left\{ s(\theta^0)'\mathscr{I}_T^{-1}(\theta^0)s(\theta^0) > c \right\},$$

as established above. If an exact value of c can be obtained, the Lagrange Multiplier test will be locally most powerful invariant for finite samples as well as asymptotically. This argument highlights the focus upon the neighborhood of the null hypothesis which is implicit in the LM procedure. King and Hillier (1980) have used this argument to establish this property in a particular case of interest where the exact critical value can be found.

7. The Lagrange Multiplier test as a diagnostic

The most familiar application of hypothesis testing is the comparison of a theory with the data. For some types of departure from the theory which might be of concern the theory may be rejected. The existence of an alternative theory is thus, very important.

A second closely related application is in the comparison of a statistical model with the data. Rarely do we know a priori the exact variables, functional forms and distribution implicit in a particular theory. Thus, there is some requirement for a specification search. At any stage in this search it may be desirable to determine whether an adequate representation of the data has been achieved. Hypothesis testing is a natural way to formulate such a question where the null hypothesis is the statistical model being used and the alternative is a more general specificiation which is being contemplated. A test statistic for this problem is called a *diagnostic* as it checks whether the data are adequately represented by the model. The exact significance of such a test is difficult to ascertain when it is one of a sequence of tests, but it should still be a sufficient statistic for the required inference and conditional on this point in the search, the size is known. In special cases of nested sequential tests, exact asymptotic significance levels can be calculated because the tests are asymptotically independent. For example see Sargan (1980) and Anderson (1971).

Frequently in applied research, the investigator will estimate several models but may not undertake comprehensive testing of the adequacy of his preferred model. Particular types of misspecification are consistently ignored. For example, the use of static models for time series data with the familiar low Durbin–Watson was tolerated for many years although now most applied workers make serial correlation corrections.

However, the next stage in generalization is to relax the "common factors" restriction implicit in serial correlation assumptions [see Hendry and Mizon (1980)] and estimate a dynamic model. Frequently, the economic implications will be very different.

This discussion argues for the presentation of a variety of diagnostics from each regression. Overfitting the model in many different directions allows the investigator to immediately assess the quality and stability of his specification.

The Lagrange Multiplier test is ideal for many of these tests as it is based upon parameters fit under the null which are therefore already available. In particular, the LM test can usually be written in terms of the residuals from the estimate under the null. Thus, it provides a way of checking the residuals for non-randomness. Each alternative considered indicates the particular type of non-randomness which might be expected.

Look for a moment at the LM test for omitted variables described in (37). The test is based upon the R^2 of the regression of the residuals on the included and potentially excluded variables. Thus, the test is based upon the squared partial correlation coefficient between the residuals and the omitted variables. This is a very intuitive way to examine residuals for non-randomness.

In the next sections, the LM test for a variety of types of misspecification will be presented. In Section 8, tests for non-spherical disturbances will be discussed while Section 9 will examine tests for misspecified mean functions including non-linearities, endogeneity, truncation and several other cases.

8. Lagrange Multiplier tests for non-spherical disturbances

A great deal of research has been directed at construction of LM tests for a variety of non-spherical disturbances. In most cases, the null hypothesis is that the disturbances are spherical; however, tests have also been developed for one type of covariance matrix against a more complicated one. In this section we will first discuss tests against various forms of heteroscedasticity as in Breusch and Pagan (1980), Engle (1982) and Godfrey (1978). Then tests against serial correlation as given by Godfrey (1978b, 1979), Breusch (1979), and Breusch and Pagan (1980) are discussed.

Test against other forms of non-spherical disturbances have also been discussed in the literature. For example, Breusch and Pagan (1980) develop a test against variance components structures and Breusch (1979) derives the tests for seemingly unrelated regression models.

8.1. Testing for heteroscedasticity

Following Breusch and Pagan (1980), let the model be specified as:

$$y_t | x_t, z_t \sim \mathrm{IN}(x_t\beta, h(z_t\alpha)) \tag{52}$$

where z_t is a $1 \times (p+1)$ vector function of x_t or other variables legitimately taken as given for this analysis. The function h is of known form with first and second derivatives and depends upon an unknown $p+1 \times 1$ vector of parameters α. The first element of z is constant with coefficient α_0 so under H_0: $\alpha_1 = \cdots = \alpha_p = 0$, the model is the classical normal regression model. The variance model includes most types of heteroscedasticity as special cases. For example, when

$$h(z_t\alpha) = \mathrm{e}^{z_t\alpha},$$

multiplicative forms are implied, while

$$h(z_t\alpha) = (z_t\alpha)^k$$

gives linear and quadratic cases for $k = 1,2$. Special case of this which might be of interest would be:

$$h(z_t, \alpha) = (\alpha_0 + \alpha_1 x_t\beta)^2,$$
$$h(z_t, \alpha) = \exp(\alpha_0 + \alpha_1 x_t\beta),$$

where the variance is related to the mean of y_t.

From applications of the formulae for the LM test given above, Breusch and Pagan derive the LM test. Letting $\theta_1 = (\alpha_1, \dots, \alpha_p)$ and $\partial h / \partial\theta_1|_{\theta_1 = 0} = \kappa z$, where κ is a scalar, the score is:

$$s(\theta^0, y) = f'z\kappa/\tilde{\sigma}^2,$$
$$\xi_{\mathrm{LM}} = \frac{T}{2} f'z(z'z)^{-1}z'f, \tag{53}$$

where $f_t = \tilde{u}_t^2/\tilde{\sigma}_t^2 - 1$, f and z are matrices with typical rows f_t and z_t and $\tilde{u}$ and $\tilde{\sigma}^2$ are the residuals and variance estimates under the null. This expression is simply one-half the explained sum of squares of a regression of f on z. As pointed out by Engle (1978), $\operatorname{plim} f'f/T = 2$ under the null and local alternatives, so an asymptotically equivalent test statistic is TR^2 from this regression. As long as z has an intercept, adding 1 to both sides and multiplying by a constant $\tilde{\sigma}^2$ will not change the R^2, thus, the statistic can be computed by regressing $\tilde{u}^2$ on z and calculating TR^2 of this regression. Koenker (1981) shows that this form is more robust to departures from normality.

The remarkable result of this test however is that κ has vanished. The test will be the same regardless of the form of h. This happens because both the score and the information matrix include only the derivative of h under H_0 and thus the overall shape of h does not matter. As far as the LM test is concerned, the alternative is:

$$h = z_t \alpha \kappa,$$

where κ is a scalar which is obviously irrelevant. This illustrates quite clearly both the strength and the weakness of local tests. One test is optimal for all h much as in the UMP case, however it seems plausible that it suffers from a failure to use the functional form of h.

Does this criticism of the LM test apply to the W and LR tests? In both cases, the parameters α must be estimated by a maximum likelihood procedure and thus the functional form of h will be important. However, the optimality of these tests is only claimed for local alternatives. For non-local alternatives the power function will generally go to one in any case and thus the shape of h is irrelevant from an asymptotic point of view. It remains possible that the finite sample non-local performance of the W and LR tests with the correct functional form for h could be superior to the LM. Against this must be set the possible computational difficulties of W and LR tests which may face convergence problems for some points in the sample space. Some Monte Carlo evidence that the LM test performs well in this type of situation is contained in Godfrey (1981).

Several special cases of this test procedure illustrate the power of the technique. Consider[1] the model $h = \exp(\alpha_0 + \alpha_1 x_t \beta)$, where H_0: $\alpha_1 = 0$. The score as calculated in (53) evaluates all parameters, including β, under the null. Thus, $x_t \tilde{\beta} = \tilde{y}_t$, the fitted values under the null. The heteroscedasticity test can be shown to have the same limiting distribution for $x_t \beta$ as for $x_t \tilde{\beta}$ and therefore it can easily be constructed as TR^2 from $\tilde{u}_t^2$ on a constant and $\tilde{y}_t$. If the model were $h = \exp(\alpha_0 + \alpha_1 (x_t \beta)^2)$ then the regression would be on a constant and $\tilde{y}_t^2$. Thus it is very easy to construct tests for a wide range of, possibly complex, alternatives.

Another interesting example is provided by the Autoregressive Conditional Heteroscedasticity (ARCH) model of Engle (1982). In this case z_t includes lagged squared residuals as well as perhaps other variables. The conditional variance is hypothesized to increase when the residuals increase. In the simplest case:

$$\begin{aligned} h &= \alpha_0 + \alpha_1 \tilde{u}_{t-1}^2 + \cdots + \alpha_p \tilde{u}_{t-p}^2 \\ &= z_t \alpha . \end{aligned}$$

This is really much like that discussed above as $\tilde{u}_{t-1} = y_{t-1} - x_{t-1} \tilde{\beta}$ and both y_{t-1}

[1] Adrian Pagan has suggested and used this model.

and x_{t-1} are legitimately taken as given in the conditional distribution. The test naturally comes out to be a regression of $\tilde{u}_t^2$ on $\tilde{u}_{t-1}^2, \ldots, \tilde{u}_{t-p}^2$ and an intercept with the statistic as TR^2 of this regression.

Once a heteroscedasticity correction has been made, it may be useful to test whether it has adequately fixed the problem. Godfrey (1979) postulates the model:

$$\sigma_t^2 = h(z_t\alpha) + g(q_t\gamma), \tag{54}$$

where $g(0) = 0$. The null hypothesis is therefore H_0: $\gamma = 0$. Under the null, estimates of $\tilde{\alpha}$ and $\tilde{u} = y_t - x_t\tilde{\beta}$ are obtained, $\tilde{\sigma}_t = h(z_t\tilde{\alpha})$ and the derivative of h at each point $z_t\tilde{\alpha}$ can be calculated as $\tilde{h}_t'$. Of course, if h is linear, this is just a constant. The test is simply again TR^2 of an auxiliary regression. In this case the regression is of:

$$\frac{\tilde{u}_t^2 - \tilde{\sigma}_t^2}{\tilde{\sigma}_t^2} \quad \text{on} \quad \frac{\tilde{h}_t' z_t}{\tilde{\sigma}_t^2} \quad \text{and} \quad \frac{q_t}{\tilde{\sigma}_t^2},$$

and the statistic will have the degrees of freedom of the number of parameters in q_t.

White (1980a) proposes a test for very general forms of heteroscedasticity. His test includes all the alternatives for which the least squares standard errors are biased. The heteroscedastic model includes all the squares and crossproducts of the data. That is, if the original model were $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, the White test would consider x_1, x_2, x_1^2, x_2^2 and $x_1 x_2$ as determinants of σ^2. The test is as usual formulated as TR^2 of a regression of u^2 on these variables plus an intercept. These are in fact just the regressors which would be used to test for random coefficients as in Breusch and Pagan (1979).

8.2. *Serial correlation*

There is now a vast literature on testing for and estimating models with serial correlation. Tests based on the LM principles are the most recent addition to the econometrician's tool kit and as they are generally very simple, attention will be confined to them.

Suppose:

$$\begin{aligned} &y_t | x_t \sim N(x_t\beta, \sigma_u^2), \\ &\alpha(L)u_t = \varepsilon, \quad u_t = y_t - x_t\beta, \quad \alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p, \end{aligned} \tag{55}$$

and ε_t is a white noise process. Then it may be of interest to test the hypothesis H_0: $\alpha_1 = \cdots = \alpha_p = 0$. Under H_0, ordinary least squares is maximum likelihood and thus the LM approach is attractive for its simplicity. An alternative formulation of (55) which shows how it fits into Crowder's framework is:

$$y_t | x_t, \psi_{t-1} \sim N(1-\alpha(L))y_t + \alpha(L)x_t\beta, \sigma_\varepsilon^2), \tag{56}$$

where ψ_{t-1} is the past information in both y and x. Thus, again under H_0 the regression simplifies to OLS but under the alternative, there are non-linear restrictions. The formulation (56) makes it clear that serial correlation can also be viewed as a restricted model relative to the general dynamic model without the non-linear restrictions. This is the common factor test which is discussed by Hendry and Mizon (1980) and Sargan (1980) and for which Engle (1979a) gives an LM test.

The likelihood function is easily written in terms of (56) and the score is simply:

$$s(y, \theta) = \frac{1}{\sigma^2} U'\tilde{u}, \tag{57}$$

where U has rows $U_t = (\tilde{u}_{t-1}, \tilde{u}_{t-2}, \dots, \tilde{u}_{t-p})$.

From the form of (57) it is clear that the LM test views U_t as an omitted set of variables from the original regression. Thus, as established more rigorously by Godfrey (1978a) and Engle (1979a), the test can be computed by regressing $\tilde{u}_t$ on x_t, U_t and testing TR^2 as a χ^2_p. The argument is essentially that because the score has the form of (31), the test will look like (38). If x_t includes no lagged dependent variables, then $\operatorname{plim} x'U/T = 0$ and the auxiliary regression will be unaffected by leaving out the x's. The test therefore is simply computed by regressing $\tilde{u}_t$ on $\tilde{u}_{t-1}, \dots, \tilde{u}_{t-p}$ and checking TR^2. For $p = 1$, this test is clearly asymptotically equivalent to the Durbin–Watson statistic.

The observation that $U'x$ will have expected value zero when x is an exogenous variable, suggests that in regression models with lagged dependent variables perhaps such products should be set to their expected value which is zero. If this is done systematically, the resulting test is Durbin's (1970) h test, at least for the first order case. Thus the h test uses the a priori structure to set some of the terms of the LM test to zero. One might expect better finite sample performance from this, however, the few Monte Carlo experiments do not show such a difference. Instead, this test performs about equally well when it exists, however, for some points in the sample space, it gives imaginary values. These apparently convey no information about the validity of the null hypothesis and are a result of the approximation of a positive definite matrix by one which is not always so. Because of this fact and the difficulty of generalizing the Durbin test for higher order serial correlation and higher order lags of dependent variables, the LM test is likely to be preferred at least for higher order problems. See Godfrey and Tremayne (1979) for further details.

It would seem attractive to construct a test against moving average disturbances. Thus suppose the model has the form:

$$\begin{aligned} &y_t|x_t \sim N(x_t\beta, \sigma_u^2), \\ &y_t - x_t\beta = u_t, \\ &u_t = \varepsilon_t - \alpha_1\varepsilon_{t-1} - \cdots - \alpha_p\varepsilon_{t-p}, \end{aligned} \tag{58}$$

where ε is again a white noise process. Then $\varepsilon_t = y_t - x_t\beta - \alpha_1\varepsilon_{t-1} - \cdots - \alpha_p\varepsilon_{t-p}$ so the log-likelihood function is proportional to:

$$L = -\sum_{t=1}^{T}\left(y_t - x_t\beta - \alpha_1\varepsilon_{t-1} - \cdots - \alpha_p\varepsilon_{t-p}\right)^2/2\sigma^2.$$

The score evaluated under the null that $\alpha_1 = \cdots = \alpha_p = 0$ is simply:

$$s(y,\tilde{\theta}) = \tilde{u}'U/\sigma^2,$$

which is identical to that in (57) for the $AR(\rho)$ model. As the null hypothesis is the same, the two tests will be the same. Again, the LM tests for different alternatives turn out to be the same test. For local alternatives, the autoregressive and moving average errors look the same and therefore one test will do for both.

When a serial correlation process has been fit for a particular model, it may still be of interest to test for higher order serial correlation. Godfrey (1978b) supposes that a (p,q) residual model has been fit and that $(p+r,q)$ is to be taken as the alternative not surprisingly, the test against $(p,q+r)$ is identical. Consider here the simplest case where $q=0$. Then the residuals under the null can be written as:

$$\begin{aligned} &\tilde{u}_t = y_t - x_t\tilde{\beta}, \\ &\tilde{\varepsilon}_t = \tilde{u}_t - \tilde{\gamma}_1\tilde{u}_{t-1} - \cdots - \tilde{\gamma}_p\tilde{u}_{t-p}. \end{aligned}$$

The test for $(p+r,0)$ or (p,r) error process can be calculated as TR^2 of the regression of $\tilde{\varepsilon}_t$ on $\tilde{x}_t, \tilde{u}_{t-1},\dots,\tilde{u}_{t-p}, \tilde{\varepsilon}_{t-1},\dots,\tilde{\varepsilon}_{t-r}$, where $\tilde{x}_t = x_t - \tilde{\gamma}_1 x_{t-1} - \cdots - \tilde{\gamma}_p x_{t-p}$. Just as in the heteroscedasticity case the regression is of transformed residuals on transformed data and the omitted variables. Here the new ingredient is the inclusion of $\tilde{u}_{t-1},\dots,\tilde{u}_{t-p}$ in the regression to account for the optimization over γ under the null.

This approach applies directly to diagnostic tests for time series models. Godfrey (1979a), Poskitt and Tremayne (1980), Hosking (1980) and Newbold

(1980) have developed and analyzed tests for a wide range of alternatives. In each case the score depends simply on the residual autocorrelations, however the tests differ from the familiar Box–Pierce–Portmanteau test in the calculation of the critical region. Consequently, the LM tests will have superior properties at least asymptotically for a finite parameterization of the alternative. If the number of parameters under test becomes large with the sample size then the tests become asymptotically equivalent. However, one might suspect that the power properties of tests against low order alternatives might make them the most suitable general purpose diagnostic tools.

When LM tests for serial correlation are derived in a simultaneous equation framework, the statistics are somewhat more complicated and in fact there are several incorrect tests in the literature. The difficulty arises over the differences in instrument lists under the null and alternative models. For a survey of this material plus presentation of several tests, see Breusch and Godfrey (1980). In the standard simultaneous equation model:

$$\begin{aligned} &Y_tB + X_t\Gamma = U_t, \\ &U_t = RU_{t-1} + \varepsilon_t, \end{aligned} \tag{59}$$

where Y and U_t are $1\times G$, X_t is $1\times K$ and R is a square $G\times G$, matrix of autoregressive coefficients, they seek to test H_0: $R=0$ both in the FIML and LIML context. They conclude that if $\tilde{U}_t$ is the set of residuals estimated under the assumption of no serial correlation, then the LM test can be approximated by any standard significance test in the augmented model:

$$Y_tB + X_t\Gamma - R\tilde{U}_{t-1} = \varepsilon_t. \tag{60}$$

Thus comparing the likelihood achieved under (59) and (60) would provide an asymptotically equivalent test to the LM test. As usual, this is just one of many computational techniques.

9. Testing the specification of the mean in several complex models

A common application of LM tests is in econometric situations where the estimation requires iterative procedures to maximize the likelihood function. In this section a variety of situations will be discussed where possibly complex misspecifications of the mean function are tested. LM tests for non-linearities, for common factor dynamics, for weak and strong exogeneity and for omitted variables in discrete choice and truncated dependent variable models are presented below. These illustrate the simplicity of LM tests in complex models and suggest countless other examples.

9.1. *Testing for non-linearities*

Frequently an empirical relationship derived from economic theory is highly non-linear. This is typically approximated by a linear regression without any test of the validity of the approximation. The LM test generally provides a simple test of such restrictions because it uses estimates only under the null hypothesis. While it is ideal for the case where the model is linear under the null and non-linear under the alternative, the procedures also greatly simplify the calculation when the null is non-linear. Three examples will be presented which show the usefulness of this set of procedures.

If the model is written as:

$$y_t | x_t \sim N\big(g(x_t, \beta), \sigma^2\big),$$

then the score under the null will have the form:

$$s(y, \tilde{\beta}) = \frac{1}{\tilde{\sigma}^2} \sum \tilde{u}_t \frac{\partial g}{\partial \beta}(x_t, \beta)|_0.$$

Thus the derivative of the non-linear relationship evaluated with parameter estimated under the null, can be considered as an omitted variable. The test would be given by the formulations in Section 5.

As an example, consider testing for a liquidity trap in the demand for money. Several studies have examined this hypothesis. Pifer (1969), White (1972) and Eisner (1971) test for a liquidity trap in logarithmic or Box–Cox functional forms while Konstas and Khouja (1969) (K–K) use a linear specification. Most studies find maximum likelihood estimates of the interest rate floor to be about 2% but they differ on whether this figure is significantly different from zero. Pifer says it is not significant, Eisner corrects his likelihood ratio test and says it is, White generalizes the form using a Box–Cox transformation and concludes that it is not different from zero. Recently Breusch and Pagan (1977a) have re-examined the Konstas and Khouja form and using a Lagrange Multiplier test, conclude that the liquidity trap is significant.

Except for minor footnotes in some of the studies, there is no mention of the serial correlation which exists in the models. In re-estimating the Konstas–Khouja model, the Durbin–Watson statistic was found to be 0.3 which is evidence of a severe problem with the specification and that the distribution of all the test statistics may be highly misleading.

The model estimated by K–K is:

$$M = \gamma Y + \beta (r - \alpha)^{-1} + \varepsilon, \tag{61}$$

where M is real money demand, Y is real GNP and r is the interest rate. Perhaps their best results are when $M1$ is used for M and the long-term government bond rate is used for r. The null hypothesis to be tested is $\alpha = 0$. The normal score is proportional to $u'z$ where z, the omitted variable, is the derivative of the right-hand side with respect to α evaluated under the null:

$$z = \left.\frac{\partial g}{\partial \alpha}\right|_0 = \frac{\beta}{r^2}.$$

Therefore, the LM test is a test of whether $1/r^2$ belongs in the regression along with Y and $1/r$.

Breusch and Pagan obtain the statistic $\xi_{\mathrm{LM}} = 11.47$ and therefore reject $\alpha = 0$. Including a constant term this becomes 5.92 which is still very significant in the X^2 table. However, correcting for serial correlation in the model under the null changes the results dramatically. A second-order autoregressive model with parameters 1.5295 and -0.5597 was required to whiten the residuals. These parameters are used in an auxiliary regression of the transformed residual on the three transformed right-hand side variables and a constant, to obtain an $R^2 = 0.01096$. This is simply GLS where the covariance parameters are assumed known. Thus, the LM statistic is $\xi_{\mathrm{LM}} = 0.515$ which is distributed as X_1^2 if the null is true. As can be seen it is very small suggesting that the liquidity trap is not significantly different from zero.

As a second example, consider testing the hypothesis that the elasticity of substitution of a production function is equal to 1 against the alternative that is constant but not unity. If y is output and x_1 and x_2 are factors of production, the model under the alternative can be written as:

$$\log y = -\frac{\alpha}{\rho}\log(\delta x_1^{-\rho} + (1-\delta)x_2^{-\rho}) + u. \tag{62}$$

If $\rho = 0$, the elasticity of substitution is one and the model becomes:

$$\log y = \alpha\delta\log x_1 + \alpha(1-\delta)\log x_2 + u.$$

To test the hypothesis $\rho = 0$, it is sufficient to calculate $\partial g/\partial\rho|_{\rho=0}$ and test whether this variable belongs in the regression. In this case

$$\left.\frac{\partial g}{\partial \rho}\right|_{\rho=0} = -\frac{\alpha}{2}\delta(1-\delta)\left(\log\frac{x_1}{x_2}\right)^2$$

which is simply the Kmenta (1967) approximation. Thus the Cobb–Douglas form can be estimated with appropriate heteroscedasticity or serial correlation and the unit elasticity assumption tested with power equal to a likelihood ratio test without ever doing a non-linear regression.

As a third example, Davidson, Hendry, Srba and Yeo (1978) estimate a consumption function for the United Kingdom which pays particular attention to the model dynamics. The equation finally chosen can be expressed as:

$$\begin{aligned}\Delta_4 c_t = {} & \beta_1\Delta_4 y_t + \beta_2\Delta_1\Delta_4 y_t + \beta_3(c_{t-4} - y_{t-4}) \\ & + \beta_4\Delta_4 D_t + \beta_5\dot{p}_t + \beta_6\Delta_1\dot{p}_t,\end{aligned} \tag{63}$$

where c, y and p are the logs of real consumption, real personal disposable income and the price level, and Δ_i is the ith difference. In a subsequent paper Hendry and Von Ungern-Sternberg (1979) argue that the income series is mismeasured in periods of inflation. The income which accrues from the holdings of financial assets should be measured by the real rate of interest rather than the nominal as is now done. There is a capital loss of $\dot{p}$ times the asset which should be netted out of income. The appropriate log income measure is $y_t^* = \log(Y_t - \alpha\dot{p}L_{t-1})$ where L is liquid assets of the personal sector and α is a scale parameter to reflect the fact that L is not all financial assets.

The previous model corresponds to $\alpha = 0$ and the argument for the respecification of the model rests on the presumption that $\alpha \neq 0$. The LM test can be easily calculated whereas the likelihood ratio and Wald tests require non-linear estimation if not respecification. The derivative of y^* with respect to α evaluated under the null is simply $-\dot{p}L_{t-1}/Y_t$. Denote this by x_t. The score is proportional to $u'z$, where $z = \tilde{\beta}_1\Delta_4 x_t + \tilde{\beta}_2\Delta_1\Delta_4 x_t - \tilde{\beta}_3 x_{t-4}$, and the betas are evaluated at their estimates under the null. This is now a one degree of freedom test and can be simply performed. The test is significant with a chi squared value of 5. As a one tailed test it is significant at the 2.5% level.

9.2. *Testing for common factor dynamics*

In a standard time series regression framework, there has been much attention given to the testing and estimation of serial correlation patterns in the disturbances. A typical model might have the form:

$$y_t = x_t\beta + u_t, \qquad \rho(L)u_t = \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{IN}(0,\sigma^2), \tag{64}$$

where $\rho(L)$ is an rth order lag polynomial and x_t is a $1\times k$ row vector which for the moment is assumed to include no lagged exogenous or endogenous variables.

Sargan (1964, 1980) and Hendry and Mizon (1978) have suggested that this is often a strong restriction on a general dynamic model. By multiplying through by $\rho(L)$ the equation can equivalently be written as:

$$\rho(L)y_t = \rho(L)x_t\beta + \varepsilon_t. \tag{65}$$

This model includes a set of non-linear parameter restrictions which essentially reduce the number of free parameters to $k+r$ instead of the full $(k+1)r$ which would be free if the restriction were not imposed. A convenient parameterization of the unrestricted alternative can be given in terms of another matrix of lag polynomials $\theta(L)$ which is a $1\times k$ row vector each element of which is an rth order lag polynomial with zero order lag equal to zero. That is $\theta(0)=0$. The unrestricted model is given by:

$$\rho(L)y_t=\rho(L)x_t\beta+\theta(L)x_t'+\varepsilon_t, \tag{66}$$

which simplifies to the serial correlation case if all elements of θ are zero. Thus, the problem can be parameterized in terms of $z=(x_{-1},\dots,x_{-r})$ as a matrix of kr omitted variables in a model estimated with GLS. The results of Section 5 apply directly. The test is simply TR^2 of $\tilde{\varepsilon}_t$ on $\tilde{\rho}(L)x_t$, z_t and $(\tilde{u}_{t-1},\dots,\tilde{u}_{t-r})$, or equivalently, on x_t, z_t $(y_{-1},\dots,y_{-r})$.

Now if x includes lags, the test must be very slightly modified. The matrix z will, in this case, include variables which are already in the model and thus the auxiliary regression will see a data set with perfect multicollinearity. The solution is to eliminate the redundant elements of z as these are not testable in any case. The test statistic will have a correspondingly reduced number of degrees of freedom.

A more complicated case occurs when it is desired to test that the correlation is of order r against the alternative that it is of order $r-1$. Here the standard test procedure breaks down. See Engle (1979a) for a discussion and some suggestions.

9.3. *Testing for exogeneity*

Tests for exogeneity are a source of controversy partly because of the variety of definitions of exogeneity implicit in the formulation of the hypotheses. In this paper the notions of weak and strong exogeneity as formulated by Engle et al. (1983) will be used in the context of linear simultaneous equation systems. In this case weak exogeneity is essentially that the equations defining weakly exogenous variables can be ignored without a loss of information. In textbook cases weakly exogenous variables are predetermined. Strong exogeneity implies, in addition, that the variables in question cannot be forecast by past values of endogenous variables which is the definition implicit in Granger (1969) "non-causality".

Consider a complete simultaneous equation system with G equations and K predetermined variables so that Y, ε, and V are $T\times G$, X is $T\times K$ and the coefficient matrices are conformable. The structural and reduced forms are:

$$YB=X\Gamma+\varepsilon, \qquad \mathrm{E}\varepsilon_t'\varepsilon_t=\Omega, \tag{67}$$

$$Y=X\Pi+V, \tag{68}$$

where ε_t are rows of ε which are independent and the x are weakly exogenous. Partitioning this set of equations into the first and the remaining $G-1$, the structure becomes:

$$y_1 - Y_2\beta = x_1\gamma + \varepsilon_1, \tag{69}$$

$$-y_1\alpha' + Y_2B_2 = X_2\Gamma_2 + \varepsilon_2, \tag{70}$$

where X_2 may be the same as X and

$$B = \begin{pmatrix} 1 & -\alpha' \\ -\beta & B_2 \end{pmatrix}, \qquad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}. \tag{71}$$

The hypothesis that Y_2 is weakly exogenous to the first equation in this full information context is simply the condition for a recursive structure:

$$H_0\colon \alpha = 0, \Omega_{12} = 0, \tag{72}$$

which is a restriction of $2G-2$ parameters.

Several variations on this basic test are implicit in the structure. If the coefficient matrix is known to be triangular, then $\alpha = 0$ is part of the maintained hypothesis and the test becomes simply a test for $\Omega_{12} = 0$. This test is also constructed below; Holly (1979) generalized the result to let the entire B matrix be assumed upper triangular and obtains a test of the diagonality of Ω and Engle (1982a) has further generalized this to block recursive systems. If some of the elements of β are known to be zero, then the testing problem remains the same. In the special case where B_2 is upper triangular between the included and excluded variables of Y_2 and the disturbances are uncorrelated with those of y_1 and the included y_2, then it is only necessary to test that the α's and Ω's of the included elements of y_2 are zero. In effect, the excluded y_2 now form a higher level block of a recursive system and the problem can be defined a priori to exclude them also from y_2. Thus without loss of generality the test in (72) can be used when some components of β take unknown values.

To test (72) with (67) maintained, first construct the normal log likelihood L, apart from some arbitrary constants:

$$L = T\log|B| - \frac{T}{2}\log|\Omega| - \tfrac{1}{2}\sum_{t=1}^{T}\varepsilon_t\Omega^{-1}\varepsilon_t'. \tag{73}$$

Partitioning this as in (71) using the identity $|\Omega| = |\Omega_{22}||\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}|$ gives:

$$\begin{aligned} L = {} & T\log|B_2| + T\log|1 - \alpha' B_2^{-1}\beta| - \frac{T}{2}\log|\Omega_{22}| \\ & - \frac{T}{2}\log|\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}| - \tfrac{1}{2}\sum_t \varepsilon_{1t}\Omega^{11}\varepsilon_{1t}' \\ & - \tfrac{1}{2}\sum_t \varepsilon_{2t}\Omega^{22}\varepsilon_{2t}' - \sum_t \varepsilon_{1t}\Omega^{12}\varepsilon_{2t}', \end{aligned} \tag{74}$$

where the superscripts on Ω indicate the partitioned inverse. Differentiating with respect to α and setting parameters to their values under the null gives the score:

$$\left.\frac{\partial L}{\partial \alpha}\right|_0 = -T\tilde{B}_2^{-1}\tilde{\beta} + \sum_t \tilde{\Omega}^{22}\tilde{U}_{2t}' y_{1t}, \tag{75}$$

where tildes represent estimates under the null and $\tilde{U}_{2t}$ is the row vector of residuals under the null. Recognizing that $\sum_t \tilde{\Omega}^{22}\tilde{U}_{2t}'\tilde{U}_{2t}/T = I$, this can be rewritten as:

$$\left.\frac{\partial L}{\partial \alpha}\right|_0 = \sum_t \tilde{\Omega}^{22}\tilde{U}_{2t}'\left(y_{1t} - \tilde{U}_{2t}\tilde{B}_2^{-1}\tilde{\beta}\right) \equiv \sum_t \tilde{\Omega}^{22}\tilde{U}_{2t}'\left(\bar{y}_{1t} + \tilde{u}_{1t}\right), \tag{76}$$

where $\bar{y}_1$ is the reduced form prediction of y_1 which is given in this case as $x_1\tilde{\gamma} + X_2\tilde{\Gamma}_2\tilde{B}_2^{-1}\tilde{\beta}$. Clearly, under the null hypothesis, the score will have expected value zero as it should. Using tensor notation this can be expressed as:

$$s_\alpha = \left(I \otimes \left(\bar{y}_1 + \tilde{u}_1\right)\right)'\left(\tilde{\Omega}_{22}^{-1} \otimes I\right)\operatorname{vec}\left(\tilde{U}_2\right), \tag{77}$$

which is in the form of omitted variables from a stacked set of regressions with covariance matrix $\tilde{\Omega}_{22}^{-1} \otimes I$. This is a GLS problem which allows calculation of a test for $\alpha = 0$ under the maintained hypothesis that $\Omega_{12} = 0$. Because of the simultaneity, the procedure in Engle (1982a) should be followed.

The other part of the test in (72) is obtained by differentiating with respect to Ω_{12} and evaluating under the null. It is not hard to show that all terms in the derivative vanish except the last. Because $\partial\Omega^{12}/\partial\Omega_{12}|_0 = -\Omega_{11}^{-1}\Omega_{22}^{-1}$ the score can be written as:

$$s_{\Omega_{12}} = \sum \tilde{u}_{1t}\tilde{\Omega}_{11}^{-1}\tilde{\Omega}_{22}^{-1}U_{2t}', \tag{78}$$

which can be written in two equivalent forms:

$$s_{\Omega_{12}} = \tilde{\Omega}_{11}^{-1}\tilde{\Omega}_{22}^{-1}\tilde{U}_2'\tilde{u}_1 \tag{79}$$

$$= \tilde{\Omega}_{11}^{-1}(I\otimes\tilde{u}_1)'(\tilde{\Omega}_{22}^{-1}\otimes I)\mathrm{vec}(\tilde{U}_2). \tag{80}$$

Either would be appropriate for testing $\Omega_{12}=0$ when $\alpha=0$ is part of the maintained hypothesis. In (79) the test would be performed in the first equation by considering U_2 as a set of $G-1$ omitted variables. In (80) the test would be performed in the other equations by stacking them and then considering $I\otimes u_1$ as the omitted set of variables. Clearly the former is easier in this case.

To perform the joint test, the two scores must be jointly tested against zero. Here (77) and (80) can easily be combined as they have just the same form. The test becomes a test for two omitted variables, $\tilde{y}_1+\tilde{u}_1$ and $\tilde{u}_1$, in each of the remaining $G-1$ equations. Equivalently, $\tilde{y}_1$ and $\tilde{u}_1$ can be considered as omitted from these equations.

Engle (1979) shows that this test can be computed as before. If the model is unidentified the test would have no power and if the model is very weakly identified, the test would be likely to have very low power.

In the special case where $G=2$, the test is especially easy to calculate because both equations can be estimated by least squares under the null. Therefore Section 5 can be applied directly.

As an example, the Michigan model of the monetary sector was examined. The equations are reported in Gardner and Hymans (1978). In this model, as in most models of the money market it is assumed that a short term interest rate can be taken as weakly exogenous in an equation for a long-term rate. However, most portfolio theories would argue that all rates are set at the same time as economic agents shift from one asset to another to clear the market.

In this example a test is constructed for the weak exogeneity of the prime rate, *RAAA*, in the 35 year government bond rate equation, *RG*35. The model can be written as:

$$\begin{aligned} RG35 &= \beta\Delta RAAA + x_1\gamma + \varepsilon_1,\\ \Delta RAAA &= \alpha RG35 + x_2\gamma + \varepsilon_2, \end{aligned} \tag{81}$$

where the estimates assume $\alpha=\sigma_{12}=0$, and the x's include a variety of presumably predetermined variables including lagged interest rates. Testing the hypothesis that $\alpha=0$ by considering *RG*35 as an omitted variable is not legitimate as it will be correlated with ε_2. If one does the test anyway, a chi-squared value of 35 is obtained.

The appropriate test of the weak exogeneity of *RG*35 is done by testing u_1 and $RG35-\tilde{\beta}\tilde{u}_2$ as omitted from the second equation where $\tilde{u}_2=\Delta RAAA-x_2\tilde{\gamma}_2$.

This test was calculated by regressing $\tilde{u}_2$ on x_2, $\tilde{u}_1$ and $RG35-\tilde{\beta}\tilde{u}_2$. The resulting $TR^2=1.25$ which is quite small, indicating that the data does not contain evidence against the hypothesis. Careful examination of x_1 and x_2 in this case shows that the identification of the model under the alternative is rather flimsy and therefore the test probably has very little power.

A second class of weak exogeneity tests can be formulated using the same analysis. These might be called limited information tests because it is assumed that there are no overidentifying restrictions available from the second block of equations. In this case equation (70) can be replaced by:

$$Y_2 = X\Pi_2 + \varepsilon_2. \tag{82}$$

Now the definition of weak exogeneity is simply that $\Omega_{12}=0$ because $\alpha=0$ imposes no restrictions on the model. This situation would be expected to occur when the second equation is only very roughly specified.

A very similar situation occurs in the case where Y_2 is possibly measured with error. Suppose Y_2^* is the true unobserved value of Y_2 but one observes $Y_2 = Y_2^* + \eta$. If the equation defining Y_2^* is:

$$Y_2^* = x_2\Gamma_2 + \varepsilon_2,$$

where the assumption that Y_2^* belongs in the first equation implies $\mathrm{E}\varepsilon_1'\varepsilon_2 = 0$, the observable equations become:

$$\begin{aligned} y_1 &= Y_2\beta + x_1\gamma + \varepsilon_1 - \eta\beta, \\ Y_2 &= x_2\Gamma_2 + \varepsilon_2 + \eta. \end{aligned} \tag{83}$$

If there is no measurement error, then the covariance matrix of η will be zero, and $\Omega_{12}=0$. This set up is now just the same as that used by Wu (1973) to test for weak exogeneity of Y_2 when it is known that $\alpha=0$.

The procedure for this test has already been developed. The two forms of the score are given in (79) and (80) and these can be used to test for the presence of U_2 in the first equation. This test is Wu's test and it is also the test derived by Hausman (1979) for this problem. By showing that these are Lagrange Multiplier tests, the asymptotic optimality of the procedures is established when the full set of x_2 is used. Neither Hausman nor Wu could establish this property.

Finally, tests for strong exogeneity can be performed. By definition, strong exogeneity requires weak exogeneity plus the non-predictability of Y_2 from past values of y_1. Partitioning x_2 in (70) into (y_1^0, x_3) where y_1^0 is a matrix with all the relevant lags of y_1, and similarly letting $\Gamma_2 = (\Gamma_{20}, \Gamma_{23})$ the hypothesis of strong exogeneity is:

$$H_0\colon \alpha = 0, \qquad \Omega_{12} = 0, \qquad \Gamma_{20} = 0. \tag{84}$$

This can clearly be jointly tested by letting u_1, $\bar{y}_1$ and y_1^0 be the omitted variables from each of the equations. Clearly the weak exogeneity and the Granger non-causality are very separate parts of the hypothesis and can be tested separately. Most often however when Granger causality is being tested on its own, the appropriate model is (82) as overidentifying restrictions are rarely available.

9.4. Discrete choice and truncated distributions

In models with discrete or truncated dependent variables, non-linear maximum likelihood estimation procedures are generally employed to estimate the parameters. The estimation techniques are sufficiently complex that model diagnostics are rarely computed and often only a limited number of specifications are tried. This is therefore another case where the LM test is useful. Two examples will be presented: a binary choice model and a self-selectivity model.

In the binary choice model, the outcome is measured by a dependent variable, y, which takes on the value 1 with probability p and 0 with probability $1-p$. For each observation these probabilities are different either because of the nature of the choice or of the chooser. Let $p_t = F(x_t\beta)$, where the function F maps the exogenous characteristics, x_t, into the unit interval. A common source of such functions are cumulative distribution functions such as the normal or the logistic. The log-likelihood of this model is given by

$$L = \sum_t \left(y_t \log p_t + (1-y_t)\log(1-p_t)\right), \qquad p_t = F(x_t\beta). \tag{85}$$

Partitioning the parameter vector and x_t vector conformably into $\beta = (\beta_1', \beta_2')'$, the hypothesis to be tested is H_0: $\beta_1 = 0$. The model has already been estimated using only x_2 as the exogenous variables and it is desired to test whether some other variables were omitted. These estimates under the null will be denoted $\tilde{\beta}_2$ which implies a set of probabilities $\tilde{p}_t$. The score and information matrix of this model are given by:

$$\frac{\partial L}{\partial \beta} = \sum_t \frac{y_t - p_t}{p_t(1-p_t)} f(x_t\beta)x_t', \tag{86}$$

$$\mathscr{I}(\beta) = \mathrm{E}\left(\frac{\partial L}{\partial \beta}\right)\left(\frac{\partial L}{\partial \beta}\right)' = \sum_t \frac{f^2(x_t\beta)}{p_t(1-p_t)} x_t'x_t, \tag{87}$$

where f is the derivative of F. Notice that the score is essentially a function of the "residuals" $y_t - p_t$. Evaluating these test statistics under the null, the LM test statistic is given by:

$$\xi_{\mathrm{LM}} = \left(\frac{\partial L}{\partial \beta}\right)' \mathscr{I}(\tilde{\beta})^{-1} \frac{\partial L}{\partial \beta}$$
$$= \tilde{u}'\tilde{x}(\tilde{x}'\tilde{x})^{-1}\tilde{x}'\tilde{u}, \tag{88}$$

where

$$\tilde{u}_t = (y_t - \tilde{p}_t)/(\tilde{p}_t(1-\tilde{p}_t))^{1/2}, \qquad \tilde{x}_t = x_t f(x_{2t}\tilde{\beta})/(\tilde{p}_t(1-\tilde{p}_t))^{1/2},$$

and

$$\tilde{u} = (\tilde{u}_1,\dots,\tilde{u}_T)', \tilde{x} = (\tilde{x}_1',\dots,\tilde{x}_T')'.$$

Because plim $\tilde{u}'\tilde{u}/T = 1$, the statistic is asymptotically equivalent to TR_0^2 of the regression of $\tilde{u}$ on $\tilde{x}$. In the special case of the logit where $p_t = 1/(1+\mathrm{e}^{-x_t\beta})$, $f = \tilde{p}_t(1-\tilde{p}_t)$ and the expressions simplify so that x_t is multiplied by $(\tilde{p}_t(1-\tilde{p}_t))^{1/2}$ rather than being divided by it. For the probit model where F is the cumulative normal, $f = \exp(x_{2_t}\tilde{\beta}_2)$ as the factor of proportionality cancels. This test is therefore extremely easy to compute based on estimates of the model under the null.

As a second example, take the self-selectivity model of Hausman and Wise (1977). The sample is truncated based upon the dependent variable. The data come from the negative income tax experiment and when the families reached a sufficiently high income level, they are dropped from the sample. Thus the model can be expressed as:

$$y|x \sim N(x\beta, \sigma^2),$$

but we only have data for $y \leq c$. Thus, the likelihood function is given as the probability density of y divided by the probability of observing this family. The log-likelihood can be expressed in terms of ϕ and Φ which are the Gaussian density and distribution functions respectively as:

$$L = \sum_t \log \phi((y_t - x_t\beta)/\sigma) - \sum_t \log \Phi((c - x_t\beta)/\sigma). \tag{89}$$

The score is:

$$\frac{\partial L}{\partial \beta} = \frac{1}{\sigma^2}\sum_t\left[y_t - x_t\beta - \sigma\phi\left(\frac{(c-x_t\beta)}{\sigma}\right)\Big/\Phi\left(\frac{(c-x_t\beta)}{\sigma}\right)\right]x_t'. \tag{90}$$

To estimate this model one sets the score to zero and solves for the parameters. Notice that this implies including another term in the regression which is the ratio of the normal density to its distribution. The inclusion of this ratio, called the Mills ratio, is a distinctive feature of much of the work of self-selectivity. The information matrix can be shown to be:

$$\mathscr{I} = \sum_t x_t'x_t\left(1+(\phi_t/\Phi_t)^2-(\phi_t/\Phi_t)(c-x_t\beta/\sigma)\right), \tag{91}$$

where $\phi_t = \phi((c-x_t\beta)/\sigma)$ and similarly for Φ_t.

To test the hypothesis H_0: $\beta_1 = 0$, denote again the estimates under the null by $\tilde\beta, \tilde\phi, \tilde\Phi$. Let $r_t^2 = 1+(\tilde\phi_t/\tilde\Phi_t)^2+(\tilde\phi_t/\tilde\Phi_t)(c-x_t\tilde\beta/\tilde\sigma)$ and define $\tilde u_t = (y_t - x_{2_t}\tilde\beta_2 + \tilde\sigma\tilde\phi_t/\tilde\Phi_t)/r_t$ and $\tilde x_t = x_t r_t$. With $\tilde u$ and $\tilde x$ being the corresponding vectors and matrices, the LM test statistic is:

$$\xi_{\mathrm{LM}} = \tilde u'\tilde x(\tilde x'\tilde x)^{-1}\tilde x'\tilde u. \tag{92}$$

As before, $\operatorname{plim}\tilde u'\tilde u/T = 1$ so an asymptotically equivalent test statistic is TR_0^2 of the regression of $\tilde u$ on $\tilde x$. Once again, the test is simply performed by a linear regression on transformed data. All of the components of this transformation such as the Mills ratio, are readily available from the preceding estimation. Thus a variety of complicated model searches and diagnostic tests can easily be carried out even in this complex maximum likelihood framework.

10. Alternative testing procedures

In this section three alternative closely related testing procedures will be briefly explained and the relationship between these methods and ones discussed in this chapter will be highlighted. The three alternatives are Neyman's (1959) $C(\alpha)$ test, Durbin's (1970) general procedure, and Hausman's (1978) specification test.

Throughout this section the parameter vector will be partitioned as $\theta' = (\theta_1', \theta_2')$ and the null hypothesis will be H_0: $\theta_1 = \theta_1^0$. Neyman's test, as exposited by Breusch and Pagan (1980), is a direct generalization of the LM test which allows consistent but inefficient estimation of the parameters θ_2 under the null. Let this estimate be $\tilde{\tilde\theta}_2$ and let $\tilde{\tilde\theta} = (\theta_1^0, \tilde{\tilde\theta}_2')'$. Expanding the score evaluated at θ around the ML estimate $\tilde\theta$ gives:

$$\frac{\partial L}{\partial\theta}(\tilde{\tilde\theta}) = \begin{pmatrix} \partial L/\partial\theta_1(\tilde\theta) \\ 0 \end{pmatrix} + \begin{pmatrix} \partial^2 L/\partial\theta_1\,\partial\theta_2'(\tilde\theta)(\tilde{\tilde\theta}_2-\tilde\theta_2) \\ \partial^2 L/\partial\theta_2\,\partial\theta_2'(\tilde\theta)(\tilde{\tilde\theta}_2-\tilde\theta_2) \end{pmatrix},$$

where $(\partial L/\partial\theta_2)(\tilde{\theta}) = 0$. Solving for the desired score:

$$\frac{\partial L}{\partial\theta_1}(\tilde{\theta}) = \frac{\partial L}{\partial\theta_1}(\tilde{\tilde{\theta}}) - \frac{\partial^2 L}{\partial\theta_1\,\partial\theta_2}(\bar{\theta})\left(\frac{\partial^2 L}{\partial\theta_2\,\partial\theta_2}(\bar{\theta})\right)^{-1}\frac{\partial L}{\partial\theta_2}(\tilde{\tilde{\theta}})$$
$$= s_1(\tilde{\tilde{\theta}}) - \mathscr{I}_{12}(\tilde{\tilde{\theta}})\mathscr{I}_{22}^{-1}(\tilde{\tilde{\theta}})s_2(\tilde{\tilde{\theta}}). \tag{93}$$

The $C(\alpha)$ test is just the LM test using (93) for the score. This adjustment can be viewed as one step of a Newton–Raphson iteration to find an efficient estimate of θ_2 based upon an initial consistent estimate. In some situations such as the one discussed in Breusch and Pagan, this results in a substantial simplification.

The Durbin (1970) procedure is also based on different estimates of the parameters. He suggests calculating the maximum likelihood estimate of θ_1 assuming $\theta_2 = \tilde{\theta}_2$, the ML estimate under the null. Letting this new estimate be $\tilde{\tilde{\theta}}_1$, the test is based upon the difference $\tilde{\tilde{\theta}}_1 - \theta_1^0$. Expanding the score with respect to θ_1 about $\tilde{\tilde{\theta}}_1$ holding $\theta_2 = \tilde{\theta}_2$ and recognizing that the first term is zero by definition of $\tilde{\tilde{\theta}}_1$ the following relationship is found:

$$\frac{\partial L}{\partial\theta_1}(\tilde{\theta}) = -\frac{\partial^2 L}{\partial\theta_1\,\partial\theta_1'}(\bar{\theta})(\tilde{\tilde{\theta}}_1 - \theta_1^0). \tag{94}$$

Because the Hessian is assumed to be non-singular, any test based upon $\tilde{\tilde{\theta}}_1 - \theta_1^0$ will have the same critical region as one based upon the score; thus the two tests are equivalent. In implementation there are of course many asymptotically equivalent forms of the tests, and it is the choice of the asymptotic form of the test which gives rise to the differences between the LM test for serial correlation and Durbin's *h* test.

The third principle is Hausman's (1978) specification test. The spirit of this test is somewhat different. The parameters of interest are not θ_1 but rather θ_2. The objective is to restrict the parameter space by setting θ_1 to some preassigned values without destroying the consistency of the estimates of θ_2. The test is based upon the difference between the efficient estimates under the null, $\tilde{\theta}_2$, and a consistent but possibly inefficient estimate under the alternative $\hat{\hat{\theta}}_2$. Hausman makes few assumptions about the properties of $\hat{\hat{\theta}}_2$; Hausman and Taylor (1980), however, modify the statement of the result somewhat to use the maximum likelihood estimate under the alternative $\hat{\theta}_2$. For the moment, this interpretation will be used here. Expanding the score around the maximum likelihood estimate and evaluating it at $\tilde{\theta}$ gives:

$$\frac{\partial L}{\partial\theta}(\tilde{\theta}) = \frac{\partial^2 L}{\partial\theta\,\partial\theta'}(\bar{\theta})(\tilde{\theta} - \hat{\theta}),$$

or

$$\begin{pmatrix} \theta_1^0 - \hat{\theta}_1 \\ \tilde{\theta}_2 - \hat{\theta}_2 \end{pmatrix} = \left(\frac{\partial^2 L}{\partial \theta \, \partial \theta'} \right)^{-1} \begin{pmatrix} \partial L / \partial \theta_1 (\tilde{\theta}) \\ 0 \end{pmatrix}. \tag{95}$$

It was shown above that asymptotically optimal tests could be based upon either the score or the difference $(\hat{\theta}_1 - \theta_1^0)$. As these are related by a non-singular transformation which asymptotically is $\mathcal{J}^{11}$, critical regions based on either statistic will be the same. Hausman's difference is based upon $\mathcal{J}^{21}$ times the score asymptotically. If this matrix is non-singular, then the tests will all be asymptotically equivalent. The dimension of $\mathcal{J}^{21}$ is $q \times p$ where p is the number of restrictions and $q = k - p$ is the number of remaining parameters. Thus a necessary condition for this test to be asymptotically equivalent is that $\min(p, q) = p$. A sufficient condition is that $\operatorname{rank}(\mathcal{J}^{21}) = p$. The equivalence requires that there be at least as many parameters unrestricted as restricted. However, parameters which are asymptotically independent of the parameters under test will not count. For example, in a classical linear regression model, the variance and any serial correlation parameters will not count in the number of unrestricted parameters. The reason for the difficulty is that the test is formulated to ignore all information in $\hat{\theta}_1 - \theta_1^0$ even though it frequently would be available from the calculation of $\hat{\theta}_2$.

Hausman and Taylor (1980) in responding to essentially this criticism from Holly (1980) point out that in the case $q < p$, the specification test can be interpreted as an asymptotically optimal test of a different hypothesis. They propose the hypothesis H_0^*: $\mathcal{J}_{22}^{-1}\mathcal{J}_{21}(\theta_1 - \theta_1^0) = 0$ or simply $\mathcal{J}_{21}(\theta_1 - \theta_1^0) = 0$. If H_0^* is true, the bias in θ_2 from restricting $\theta_1 = \theta_1^0$ would asymptotically be zero. The hypothesis H_0^* is explicitly a consistency hypothesis. The Hausman test is one of many asymptotically equivalent ways to test this hypothesis. In fact, the same Wald, LR and LM tests are available as pointed out by Riess (1982). The investigator must however decide which hypothesis he wishes to test, H_0 or H_0^*.

In answering the question of which hypothesis is relevant, it is important to ask why the test is being undertaken in the first place. As the parameters of interest are θ_2, the main purpose of the test is to find a more parsimonious specification, and the advantage of a parsimonious specification is that more efficient estimates of the parameters of interest can be obtained. Thus if consistency were the only concern of the investigator, he would not bother to restrict the model at all. The objective is therefore to improve the efficiency of the estimation by testing and then imposing some restrictions. These restrictions ought, however, to be grounded in an economic hypothesis rather than purely data based as is likely to be the case for H_0^* which simply asserts that the true parameters lie in the column null space of $\mathcal{J}_{21}$.

Finally, if an inefficient estimator $\hat{\theta}$ is used in the test, it is unlikely that the results will be as strong as described above. Except in special cases, one would expect the test based upon the MLE to be more powerful than that based upon an inefficient estimator. However, this is an easy problem to correct. Starting from the inefficient estimate, one step of a Newton–Raphson type algorithm will produce asymptotically efficient parameter estimates.

11. Non-standard situations

While many non-standard situations may arise in practice, two will be discussed here. The first considers the properties of the Wald, LM and LR tests when the likelihood function is misspecified. The second looks at the case where the information matrix is singular under the null.

White (1982) and Domowitz and White (1982) have recently examined the problem of inference in maximum likelihood situations where the wrong likelihood has been maximized. These quasi-maximum likelihood estimates may well be consistent, however the standard errors derived from the information matrix are not correct. For example, the disturbances may be assumed to be normally distributed when in fact they are double exponentials. White has proposed generalizations of the Wald and LM test principles which do have the right size and which are asymptotically powerful when the density is correctly assumed. These are derived from the fact that the two expressions for the information matrix are no longer equivalent for QML estimates. The expectation of the outer product of the scores does not equal minus the expectation of the Hessian. Letting L_t be the log-likelihood of the tth observation, White constructs the matrices:

$$A=\frac{1}{T}\frac{\partial^2 L}{\partial\theta\,\partial\theta'};\qquad B=\frac{1}{T}\sum_t\frac{\partial L_t}{\partial\theta}\left(\frac{\partial L_t}{\partial\theta}\right)'\quad\text{and}\quad C=A^{-1}BA^{-1}.$$

Then the "quasi-scores", measured as the derivative of the possibly incorrect likelihood function evaluated under the null, will have a limiting distribution based upon these matrices when the null is true. Letting A^{11} be the first block of the partitioned inverse of A, the limiting covariance of the quasi score is $(A^{11}C_{11}^{-1}A^{11})^{-1}$ so the quasi-LM test is simply:

$$\xi_{\text{LM}}=s'A^{11}C_{11}^{-1}A^{11}s.$$

Notice that if the distribution is correct, then $A=-B$ so that $C=A^{-1}$ and the whole term becomes simply A^{11} as usual. Thus the use of the quasi-LM statistic corrects the size of the test when the distribution is false but gives the asymptotically optimal test when it is true. Except for possible finite sample and computational costs, it appears to be a sensible procedure. Exactly the same correction is made to the Wald test to obtain a quasi Wald test. Because it is the divergence between A and B which creates the situation, White proposes an omnibus test for differences between A and B.

In some situations, an alternative to this approach would be to test for normality directly as well as for other departures from the specification. Jarque and Bera (1980, 1982) propose such a test by taking the Pearson density as the alternative and simultaneously testing for serial correlation, functional form misspecification and heteroscedasticity. This joint test decomposes into independent LM tests because of the block diagonality of the information matrix for this problem.

A second non-standard situation which occurs periodically in practice is when some of the parameters are estimable only when the null hypothesis is false. That is, the information matrix under the null is singular. Two simple examples with rather different conclusions are:

$$y|x_1;x_2 \sim N(\alpha\beta x_1 + \beta x_2, \sigma^2), \qquad H_0\colon \beta = 0,$$

$$y|x \sim N(\beta x^{\alpha}, \sigma^2), \qquad H_0\colon \beta = 0.$$

In both cases, the likelihood function can be maximized under both the null and alternative, but the limiting distribution of the likelihood ratio statistic is not clear. Furthermore, conventional Wald and LM tests also have difficulties—the LM will have a parameter which is unidentified under the null which appears in the score, and the Wald will have an unknown limiting distribution. In the first example, it is easy to see that by reparameterizing the model, the null hypothesis becomes a two degree of freedom standard test. In the second example, however, there is no simple solution. Unless the parameter α is given a priori, the tests will have the above-mentioned problems. A solution proposed by Davies (1977) is to obtain the LM test statistic for each value of the unidentified parameter and then base the test on the maximum of these. Any one of these would be chi squared with one degree of freedom, however, the maximum of a set of dependent chi squares would not be chi squared in general. Davies finds a bound for the distribution which gives a test with size less than or equal to the nominal value.

As an example of this, Watson (1982) considers the problem of testing whether a regression coefficient is constant or whether it follows a first order autoregressive process. The model can be expressed as:

$$y_t = x_t\beta_t + z_t\gamma + \varepsilon_t,$$

$$\beta_t = \rho\beta_{t-1} + \eta_t,$$

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix}\right).$$

The null hypothesis is that $\sigma_{\eta}^2 = 0$; this however makes the parameter ρ unidentifiable. The test is constructed by first searching over the possible values of ρ to find the maximum LM test statistic, and then finding the limiting distribution of the test to determine the critical value. A Monte Carlo evaluation of the test showed it to work reasonably well except for values of ρ close to unity when the limiting distribution was well approximated only for quite large samples.

Several other applications of this result occur in econometrics. In factor analytical models, the number of parameters varies with the number of factors so testing the number of factors may involve such a problem. Testing a series for white noise against an $AR(1)$ plus noise again leads to this problem as the parameter in the autoregression is not identified under the null. A closely related problem occurred in testing for common factor dynamics as shown in Engle (1979a). Several others could be illustrated.

12. Conclusion

In a maximum likelihood framework, the Wald, Likelihood Ratio and Lagrange Multiplier tests are a natural trio. They all share the property of being asymptotically locally most powerful invariant tests and in fact all are asymptotically equivalent. However, in practice there are substantial differences in the way the tests look at particular models. Frequently when one is very complex, another will be much simpler. Furthermore, this formulation guides the intuition as to what is testable and how best to formulate a model in order to test it. In terms of forming diagnostic tests, the LM test is frequently computationally convenient as many of the test statistics are already available from the estimation of the null.

The application of these test principles and particularly the LM principle to a wide range of econometric problems is a natural development of the field and it is a development which is proceeding at a very rapid pace. Soon, most of the interesting cases will have been touched in theoretical papers, however, applied work is just beginning to incorporate these techniques and there is a rich future there.

References

Aitcheson, J. and S. D. Silvey (1958), "Maximum Likelihood Estimation of Parameters Subject to Restraints", *Annals of Mathematical Statistics*, 29:813–828.

Anderson, T. W. (1971), *The Statistical Analysis of Time Series*. New York: John Wiley and Sons.

Bera, A. K. and C. M. Jarque (1982), "Model Specification Tests: A Simultaneous Approach", *Journal of Econometrics*, 20:59–82.

Berndt, E. R. and N. E. Savin (1977), "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model", *Econometrica*, 45:1263–1278.

Breusch, T. S. (1978), "Testing for Autocorrelation in Dynamic Linear Models", *Australian Economic Papers*, 17:334–355.

Breusch, T. S. and A. R. Pagan (1979), "A Simple Test for Heteroskedasticity and Random Coefficient Variation", *Econometrica*, 47:1287–1294.

Breusch, T. S. (1979), "Conflict Among Criteria for Testing Hypotheses: Extensions and Comments", *Econometrica*, 47:203–207.

Breusch, T. S. and L. G. Godfrey (1980), "A Review of Recent Work on Testing for Autocorrelation in Dynamic Economic Models", Discussion Paper #8017, University of Southampton.

Breusch, T. S. and A. R. Pagan (1980), "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics", *Review of Economic Studies*, 47:239–254.

Cox, D. R. and D. V. Hinckley (1974), *Theoretical Statistics*. London: Chapman and Hall.

Crowder, M. J. (1976), "Maximum Likelihood Estimation for Dependent Observations", *Journal of the Royal Statistical Society, Series B*, 45–53.

Davidson, J. E. H., Hendry, D. F., Srba, F., and S. Yeo (1978), "Econometric Modelling of the Aggregate Time-Series Relationship Between Consumers' Expenditure and Income in the United Kingdom", *Economic Journal*, 88:661–692.

Davies, R. B. (1977), "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative", *Biometrika*, 64:247–254.

Domowitz, I. and H. White (1982), "Misspecified Models with Dependent Observations", *Journal of Econometrics*, 20:35–58.

Durbin, J. (1970), "Testing for Serial Correlation in Least Squares Regression When Some of the Regressors are Lagged Dependent Variables", *Econometrica*, 38:410–421.

Eisner, R. (1971), "Non-linear Estimates of the Liquidity Trap", *Econometrica*, 39:861–864.

Engle, R. F. (1979), "Estimation of the Price Elasticity of Demand Facing Metropolitan Producers", *Journal of Urban Economics*, 6:42–64.

Engle, R. F. (1982), "Autoregression Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation", *Econometrica*, 50:987–1007.

Engle, R. F. (1979a), "A General Approach to the Construction of Model Diagnostics Based on the Lagrange Multiplier Principle", U.C.S.D. Discussion Paper 79-43.

Engle, R. F. (1982a), "A General Approach to Lagrange Multiplier Model Diagnostics", *Journal of Econometrics*, 20:83–104.

Engle, R. F. (1980), "Hypothesis Testing in Spectral Regression: the Lagrange Multiplier as a Regression Diagnostic", in: Kmenta and Ramsey, eds., *Criteria for Evaluation of Econometric Models*. New York: Academic Press.

Engle, R. F., D. F. Hendry, and J. F. Richard (1983), "Exogeneity", *Econometrica*, 50:227–304.

Evans, G. B. A. and N. E. Savin (1982), "Conflict Among the Criteria Revisited; The W, LR and LM tests", *Econometrica*, 50:737–748.

Ferguson, T. S. (1967), *Mathematical Statistics*. New York: Academic Press.

Godfrey, L. G. (1978), "Testing for Multiplicative Heteroskedasticity", *Journal of Econometrics*, 8:227–236.

Godfrey, L. G. (1978a), "Testing Against general Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables", *Econometrica*, 46:1293–1302.

Godfrey, L. G. (1978b), "Testing for Higher Order Serial Correlation in Regression Equations when the Regressors Include Lagged Dependent Variables", *Econometrica*, 46:1303–1310.

Godfrey, L. G. (1979), "A Diagnostic Check on the Variance Model in Regression Equations with Heteroskedastic Disturbances", unpublished manuscript, University of York.

Godfrey, L. G. (1979a), "Testing the Adequacy of a Time Series Model", *Biometrika*, 66:67–72.

Godfrey, L. G. and A. R. Tremayne (1979), "A Note on Testing for Fourth Order Autocorrelation in Dynamic Quarterly Regression Equations", unpublished manuscript, University of York.

Godfrey, L. G. (1980), "On the Invariance of the Lagrange Multiplier Test with Respect to Certain Changes in the Alternative Hypothesis", *Econometrica*, 49:1443–1456.

Hausman, J. (1978), "Specification Tests in Econometrics", *Econometrica*, 46:1251–1272.

Hausman, J. and D. Wise (1977), "Social Experimentation Truncated Distributions, and Efficient Estimation", *Econometrica*, 45:319–339.

Hausman, J. and W. Taylor (1980), "Comparing Specification Tests and Classical Tests", unpublished manuscript.

Hendry, D. F. and T. von Ungern-Sternberg (1979), "Liquidity and Inflation Effects on Consumers'

Expenditure", in: Angus Deaton, ed., *Festschrift for Richard Stone*. Cambridge: Cambridge University Press.

Hendry, D. F. and G. Mizon (1980), "An Empricial Application and Monte Carlo Analysis of Tests of Dynamic Specification", *Review of Economic Studies*, 47:21–46.

Holly, A. (1982), "A Remark on Hausman's Specification Test," *Econometrica*, v. 50: 749–759.

Hosking, J. R. M. (1980), "Lagrange Multiplier Tests of Time Series Models", *Journal of the Royal Statistical Society B*, 42:170–181.

Jarque, C. and A. K. Bera (1980), "Efficient Tests for Normality, Homoscedasticity, and Serial Independence of Regression Residuals", *Economics Letters*, 6:255–259.

King, M. L. and G. H. Hillier (1980), "A Small Sample Power Property of the Lagrange Multiplier Test", Discussion Paper, Monash University.

Kmenta, J. (1967), "On Estimation of the CES Production Function", *International Economic Review*, 8:180–189.

Koenker, R. (1981), "A Note on Studentizing a Test for Heteroscedasticity", *Journal of Econometrics*, 17:107–112.

Konstas, P. and M. Khouja (1969), "The Keynesian Demand-for-Money Function: Another Look and Some Additional", *Journal of Money Credit and Banking*, 1:765–777.

Lehmann, E. L. (1959), *Testing Statistical Hypotheses*. New York: John Wiley and Sons.

Neyman, J. (1959), "Optimal Asymptotic Tests of Composite Statistical Hypotheses", in (U. Grenander, ed.) *Probability and Statistics*. Stockholm: Almquist and Wiksell, pp. 213–234.

Newbold, P. (1980), "The Equivalence of Two Tests of Time Series Model Adequacy", *Biometrica*, 67:463–465.

Pifer, H. (1969), "A Non-linear Maximum Likelihood Estimate of the Liquidity trap," *Econometrica*, 37:324–332.

Poskitt, D.S. and A.P. Tremayne (1980), "Testing the Specification of a Fitted ARMA Model", *Biometrica*, 67:359–363.

Rao, C. R. (1948), "Large Sample Tests of Statistical Hypothese Concerning Several Parameters with Application to Problems of Estimation", *Proceedings of the Cambridge Philosophical Society*, 44:50–57.

Reiss, P. (1982), "Alternative Interpretations of Hausman's *m* Test", manuscript Yale University.

Rothenberg, T. J. (1980), "Comparing Alternative Asymptotically Equivalent Tests", invited paper presented at World Congress of the Econometric Society, Aix-en-Provence, 1980.

Sargan, J. D. (1964), "Wages and Prices in the United Kingdom: A Study in Econometric Methodology", in (P.E. Hart, G. Mills, J.K. Whitaker, eds.) *Econometric Analysis for National Economic Planning*. London: Butterworths, 1964.

Sargan, J. D. (1980), "Some Tests of Dynamic Specification for a Single Equation", *Econometrica*, 48:879–897.

Savin, N. E. (1976), "Conflicts Among Testing Procedures in a Linear Regression Model with Autoregressive Disturbances", *Econometrica*, 44:1303–1313.

Silvey, D. S. (1959), "The Lagrangean Multiplier Test", *Annals of Mathematical Statistics*, 30:389–407.

Wald, A. (1943), "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large", *Transactions of the American Mathematical Society*, 54:426–482.

Watson, M. (1982). "A Test for Regression Coefficient Stability When a Parameter is Identified Only Under the Alternative", Harvard Discussion Paper 906.

White, H. (1980), "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica*, 48:817–838.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, 50:1–26.

White, K. (1972), "Estimation of the Liquidity Trap With a Generalized Functional Form", *Econometrica*, 40:193–199.

Wilks, S. S. (1938), "The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses", *Annals of Mathematical Statistics*, 9:60–62.