SOME ASPECTS OF STATISTICAL INFERENCE FOR ECONOMETRICS

By

TREVOR STANLEY BREUSCH

DECLARATION

The contents of this thesis are my own work, except where otherwise indicated.

T.S. Breusch
ACKNOWLEDGEMENTS

My greatest debt is to my thesis supervisor, Dr A.R. Pagan. For his guidance and criticism, and especially for his enthusiasm, I am most grateful.

Many others in the Departments of Statistics and Economics, Faculty of Economics, A.N.U., contributed in some way to the environment in which I have worked. In particular, I would like to thank Dr D.F. Nicholls who supervised me during Dr Pagan's absence on leave, and Professor R.D. Terrell whose door was always open for advice on matters great and small.

My thanks also go to Ms H. Patrikka who typed the manuscript with skill and care.
ABSTRACT

This thesis is concerned with examining relationships among the various asymptotic hypothesis testing principles in econometric settings and with developing applications of the Lagrange multiplier (LM) procedure to econometric problems. For a wide range of hypothesis testing situations, particularly those associated with detecting mis-specification errors in regression models, it is argued that the LM method is most useful. The LM test, which is asymptotically equivalent to the likelihood ratio test in regular problems, is frequently less demanding computationally than other procedures that might be applied in the same circumstances. In addition, the LM statistic sometimes corresponds to a criterion which is familiar to the econometrician but which has been previously motivated by other considerations. The LM testing principle provides a convenient framework in which such existing tests can be extended and new tests can be developed.

Chapter 1 sketches the theoretical setting that is applicable to many statistical problems in econometrics and highlights a number of aspects of the various testing principles, for reference in later chapters. Tests of coefficient restrictions in linear regression models are considered in Chapter 2, including an examination of a systematic numerical inequality relationship among the criteria. Chapter 3 is concerned with the LM test in its various guises and with applicability of the LM method to diverse econometric situations. Specific applications are considered in greater detail in Chapters 4 through 6: in Chapter 4 the LM method is applied to testing for autocorrelation in dynamic single equation linear models; in Chapter 5 the ideas of the preceding chapter are extended to simultaneous equations systems, and in Chapter 6 a test against a wide class of heteroscedastic disturbance formulations is
developed. Since the theoretical properties of the LM test derive mainly from asymptotic considerations, questions regarding the validity of asymptotic results to practical situations with finite sample sizes remain open. A Monte Carlo simulation study, comparing the LM test for heteroscedasticity with other asymptotically equivalent tests, is presented in Chapter 7.
# Table of Contents

## Declaration

Page (ii)

## Acknowledgements

Page (iii)

## Abstract

Page (iv)

## Chapter 1. Inference Principles for Econometrics

1.1 Introduction 1
1.2 The Inference Framework 4
1.3 The Testing Criteria 9
1.4 Simplifications for Some Special Cases 13
1.5 Relative Power Properties 15
1.6 Relative Computational Requirements 19
1.7 Problems with Parameters on Boundaries 21
1.8 Singular Information Matrices 23
1.9 Two-Step Estimators 25

## Chapter 2. Testing Coefficient Constraints in Linear Regression

2.1 Introduction 29
2.2 Notation and Derivation of the Statistics 30
2.3 The Inequality Relationship 35
2.4 Scalar Covariance Matrix 38
2.5 Nested Hypotheses 40
2.6 Nonlinearities 42
2.7 Other Extensions 44
2.8 Significance of the Inequality 46

## Chapter 3. The Lagrange Multiplier Test and Its Application to Model Specification

3.1 Introduction 49
3.2 Forms of the LM Statistic 51
3.3 The C(a) Statistic 55
3.4 Nonlinear Regression Models 57
3.5 Testing for Nonspherical Disturbances 65
3.6 Testing Diagonality of the Covariance Matrix in SUR Systems 71
3.7 The Chow Test 73
3.8 Further Aspects of the LM Test 76

## Chapter 4. Testing for Autocorrelation in Dynamic Linear Models

4.1 Introduction 79
4.2 Durbin's Methodology and the LM Test 81
4.3 Autoregressive Disturbances 85
4.4 Moving Average Disturbances 92
4.5 Some Special Cases 96
4.6 Alternative Forms of the LM Statistic 101
<table>
<thead>
<tr>
<th>CHAPTER 5. TESTING FOR AUTOREGRESSIVE DISTURBANCES IN DYNAMIC SIMULTANEOUS EQUATIONS SYSTEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Introduction 108</td>
</tr>
<tr>
<td>5.2 The Model and the LM Test 110</td>
</tr>
<tr>
<td>5.3 Comments on Previous Formulations 116</td>
</tr>
<tr>
<td>5.4 Alternative Computational Methods 119</td>
</tr>
<tr>
<td>5.5 Limited Information 124</td>
</tr>
<tr>
<td>5.6 Concluding Comments 134</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 6. TESTING FOR HETEROSEDASTICITY INCLUDING RANDOM COEFFICIENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Introduction 136</td>
</tr>
<tr>
<td>6.2 The Model with Heteroscedastic Disturbances 137</td>
</tr>
<tr>
<td>6.3 Difficulties with Maximum Likelihood Estimation of Certain Heteroscedastic Schemes 140</td>
</tr>
<tr>
<td>6.4 Derivation of the LM Statistic for Heteroscedasticity 143</td>
</tr>
<tr>
<td>6.5 Finite-Sample Properties 147</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 7. A SIMULATION STUDY COMPARING W, LR AND LM TESTS FOR HETEROSEDASTICITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1 Introduction 159</td>
</tr>
<tr>
<td>7.2 The Model and Estimation Procedure 160</td>
</tr>
<tr>
<td>7.3 Experimental Design 166</td>
</tr>
<tr>
<td>7.4 Monte Carlo Techniques 169</td>
</tr>
<tr>
<td>7.5 Experimental Results 172</td>
</tr>
</tbody>
</table>

APPENDIX A. Experimental Results (Tables) 180
APPENDIX B. Exogenous Variables Data 185
APPENDIX C. Invariance of the Test Statistics to the Values of Certain Parameters 187

<table>
<thead>
<tr>
<th>CHAPTER 8. CONCLUSIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>REFERENCES 195</td>
</tr>
</tbody>
</table>
CHAPTER 1
INFERENCE PRINCIPLES FOR ECONOMETRICS

1.1 Introduction

Research into statistical methodology for econometric modelling has tended to concentrate upon methods for estimation of the unknown parameters in theoretical specifications. In contrast, relatively little attention has been given to devising tools for discriminating between alternative specifications. The range of methods generally applied to the companion problem in statistical inference, that of testing hypotheses about the parameters, is somewhat limited compared with the degree of sophistication of estimators that are now available. Familiar "t-ratios" for testing whether a parameter can be set to zero and, in certain situations, the likelihood ratio criterion would be foremost in the econometrician's kit of testing tools.

Of course, many special tests have been proposed, particularly for detecting misspecification errors such as autocorrelation and heteroscedasticity in regression models, where re-estimating a more general specification and applying the likelihood ratio test might be computationally unattractive. However, these tests are usually one-off developments for the particular problem in hand. Sometimes they are also ad hoc measures, obtained without reference to any general hypothesis testing principle that would ensure that the criterion has at least some desirable properties.

One reason for the emphasis upon estimation methods in econometrics is that the models which are used tend to be complicated statistical constructions. Although the basic statistical framework of econometric
modelling is the linear regression relationship, the process of research has suggested many modifications (e.g. nonlinearities, dynamic relationships, simultaneous equations) which complicate the problems of statistical inference. Much effort has therefore been expended in devising appropriate estimators and analyzing their properties.

Another aspect of the complicated statistical nature of econometric models is that it is usually impractical to calculate exact finite-sample distributional properties of estimators and test statistics. Therefore analysis of statistical questions is typically restricted to asymptotic theory, i.e. limiting behaviour as the sample size tends to infinity. From asymptotic theory, there are three general principles for constructing tests of parametric hypotheses, each of which is closely related to the estimation method of maximum likelihood. With an hypothesized relationship between the parameters to be tested, there are two estimation situations to be distinguished: one is the restricted case where maximization is constrained over the subset of the parameter space in which the parameters satisfy the hypothesis, the other is the unrestricted case where the hypothesis is ignored in estimation. To assess agreement of sample evidence with the hypothesis, comparisons between the outcomes of the two estimation problems can be made in terms of the relative locations of the maxima, the relative suprema of the likelihood that are attained, or the extent to which constrained estimates fail to satisfy the conditions for full unrestricted maximization. In brief:

(i) the Wald test \( W \) considers whether unrestricted estimates satisfy the hypothesized relationship between the parameters;

(ii) the likelihood ratio procedure \( LR \) compares the values of maximized likelihoods, with and without the hypothesis imposed; and
(iii) the Lagrange multiplier method (LM) tests the effect of imposing the hypothesis upon the first-order conditions for a maximum of the unrestricted likelihood.

In the econometrics literature, most attention seems to have been centred on the first two of these three principles. Usually "t-tests" would rely upon the W principle for their large-sample validity, while there have been a number of papers advocating and illustrating use of the LR procedure.

This thesis is concerned with examining relationships among the three asymptotic testing principles in econometric settings and with developing applications of the LM testing procedure to econometric problems. A continuing theme is the applicability of the LM method to a wide range of hypothesis testing problems, particularly those associated with testing for misspecification errors in regression models. In many situations, it will be noted that the LM procedure is attractive for its computational simplicity and economy relative to W or LR tests, thereby providing researchers with a simple technique for assessing the adequacy of their specifications. Additionally, it will be observed that the criterion which is given by the LM principle for some problems coincides with a statistic that has earlier been developed for the same situation by a different method. Usage of the LM testing principle enables relationships between existing tests to be established and provides a framework in which new tests can be developed.

In the remainder of this chapter, the general asymptotic theory upon which the three testing methods are based is sketched, the test statistics are defined in relation to the two estimation situations, and general relationships among the three criteria are discussed. Attention is also
drawn to a number of nonstandard conditions which occur frequently in econometric problems and which may require modification of the testing approaches. The discussion here is in some detail but is rather informal from a mathematical point of view; more formal treatments of the method of maximum likelihood and associated test statistics are given by, inter alia, Silvey (1959) and Le Cam (1956, 1970). A good survey, including a number of more recent developments but unfortunately concentrating upon problems with a single parameter, may be found in Cox and Hinkley (1974, Ch. 9).

1.2 The Inference Framework

The general framework to be considered is that of a likelihood, $L(\theta)$, which is formed from $n$ observations and which is known except for $s$ unknown parameters, $\theta = (\theta_1, \ldots, \theta_s)'$. Sufficient regularity of the likelihood is assumed for the usual results of asymptotic maximum likelihood theory to be obtained.

To be tested is a null hypothesis which asserts $p < s$ restrictions on $\theta$, specified as

$$H_0: \phi(\theta) = 0$$

where $\phi(\theta)$ is a $p$-vector function of the $s$-vector $\theta$. No redundancies within the set of restrictions will be allowed, so that the $s \times p$ matrix

---

1 For notational clarity, the realizations of the random variables upon which the likelihood is conditioned are not shown explicitly as arguments of $L(\cdot)$. In Appendix C to Chapter 7 where such a distinction needs to be made, a slightly different notation is used.

2 See, for example, Cox and Hinkley (1974, p.281). Deviations from the standard conditions that arise in certain econometric applications are noted in §1.7 referring to parameter values lying on the boundary of the parameter space and in §1.8 for situations in which the parameters of the unrestricted model are unidentified.
\[ F(\theta) = \frac{\partial \phi}{\partial \theta} (\theta) \]

will be taken as having full column rank of \( p \). The alternative hypothesis will generally be \( \phi(\theta) \neq 0 \), although some particular cases in which one-sided alternatives may be appropriate will be noted in the sequel.

Writing the (natural) logarithm of the likelihood as \( \ell(\theta) = \log L(\theta) \), the score vector is

\[ d(\theta) = \frac{\partial \ell}{\partial \theta} (\theta) \]

where \( E[d(\theta)] = 0 \) when the argument \( \theta \) is the true parameter value. The information matrix,

\[ I(\theta) = E[d(\theta)(d(\theta))'] = E \left[ \frac{\partial^2 \ell}{\partial \theta \partial \theta'} (\theta) \right] \]

is assumed to be positive definite, except in §1.8 where treatment of a singular information matrix is considered. Under fairly general conditions that are almost always reasonable assumptions for econometric problems, a central limit theorem will apply to the score vector such that

\[ I^{-\frac{1}{2}} d(\theta) \overset{D}{\rightarrow} N(0, I_s) \]

(1)

where \( I^{\frac{1}{2}} \) is the unique positive definite square root of \( I \). Also, a law of large numbers will apply to the second derivative matrix of the log-likelihood so that, while the diagonal elements of \( I \) become infinitely large as \( n \to \infty \),

\[ - I^{-1} \left[ \frac{\partial^2 \ell}{\partial \theta \partial \theta'} (\theta) \right] \overset{P}{\rightarrow} I_s . \]

(2)

Here "\( D \)" denotes convergence in distribution and "\( P \)" indicates convergence in probability.

Note that (1) and (2) are somewhat more general than is typically
established (or assumed) in econometric applications. The usual econometric situation has the likelihood conditioned upon vectors of exogenous regressor variables \( x_t \) for \( t = 1, \ldots, n \), about which it is assumed that \( n^{-1} \sum_{t=1}^{n} x_t x'_t \) converges to a finite matrix as \( n \to \infty \). Then (1) can be replaced by the more familiar:

\[
n^{-\frac{1}{2}} d(\hat{\theta}) \overset{D}{\to} N(0, \lim_{n \to \infty} n^{-1} I) .
\]

However, the more general setting adopted here allows a wider class of exogenous variables, e.g. those satisfying Grenander's conditions [defined in Hannan (1970, p.77)]. Asymptotic results (1) and (2) can also be established for many cases where the score vector is a linear combination of independently but not identically distributed random variables [e.g. Eicker (1966)] and for likelihoods formed from dependently distributed random variables [e.g. Crowder (1976)].

Denote by \( \hat{\theta} \) the vector of estimates given by maximizing \( L(\theta) \), or equivalently \( \ell(\theta) \), when the hypothesized restrictions are ignored. Consistency of \( \hat{\theta} \) for \( \theta \) and asymptotic normality of \( (\hat{\theta} - \theta) \), when suitably normed, will generally follow from the conditions given above. The asymptotic distribution of \( I^{1/2}(\hat{\theta} - \theta) \) can then be found by a local linearization of the first-order conditions for a maximum of \( \ell(\theta) \) at \( \theta = \hat{\theta} \), viz.

\[
d(\hat{\theta}) = 0. \quad \text{By the mean value theorem,}
\]

\[
d(\hat{\theta}) = 0 = d(\theta) + \left[ \frac{\partial^2 \ell}{\partial \theta \partial \theta'}(\theta^*) \right] (\hat{\theta} - \theta)
\]

where \( |\theta^*_j - \theta_j*| < |\theta^*_j - \hat{\theta}_j| \) for \( j = 1, \ldots, s \) so that \( \hat{\theta} \overset{P}{\to} \theta \) implies \( \theta^* \overset{P}{\to} \theta \). Then from (2),

\[
-I^{1/2}[\frac{\partial^2 \ell}{\partial \theta \partial \theta'}(\theta^*)] I^{1/2} \overset{P}{\to} I_s
\]

so that premultiplying through (3) by \( I^{-1/2} \) and rearranging gives
\[ I_{\theta}^\frac{1}{2}(\theta - \theta_0) = \left\{ I_{\theta}^{-\frac{1}{2}}[\theta^2 / \partial \theta \partial \theta'](\theta^*) \right\} I_{\theta}^{-\frac{1}{2}} \left[ \begin{array}{c} \frac{I_{\theta}^{1/2}}{d(\theta)} + o_p(1) \\ N(0, \Sigma) \end{array} \right] \]  

In (5) the order of magnitude of the approximation term comes from the convergence in distribution of \( I_{\theta}^{-\frac{1}{2}} d(\theta) \) and the probabilistic convergence in (4). Expression (6) gives the asymptotic distribution of the unrestricted maximum likelihood estimator.

The constrained estimation problem requires the \( \theta \) value which maximizes \( \ell(\theta) \) while satisfying the hypothesis \( \phi(\theta) = 0 \). Following Aitchison and Silvey (1958), the Lagrangian function

\[ \psi(\theta, \lambda) = \ell(\theta) + \lambda^T \phi(\theta) \]

is formed with \( \lambda \) a \( p \)-vector of unknown Lagrange multipliers. For a maximum of \( \psi(\theta, \lambda) \) at \( \theta = \tilde{\theta} \) and \( \lambda = \tilde{\lambda} \), the first-order conditions are

\[ \frac{\partial \psi}{\partial \theta}(\tilde{\theta}, \tilde{\lambda}) = d(\tilde{\theta}) + [F(\tilde{\theta})] \tilde{\lambda} = 0 \]  

(7)

\[ \frac{\partial \psi}{\partial \lambda}(\tilde{\theta}) = \phi(\tilde{\theta}) = 0 . \]  

(8)

Expanding \( d(\tilde{\theta}) \) and \( \phi(\tilde{\theta}) \) gives

\[ d(\tilde{\theta}) = d(\theta) + [\theta^2 / \partial \theta \partial \theta'](\theta^*)](\tilde{\theta} - \theta) \]  

(9)

\[ \phi(\tilde{\theta}) = \phi(\theta) + [F(\theta^*)]'(\tilde{\theta} - \theta) \]  

(10)

The notation \( o_p(\cdot) \) and \( O_p(\cdot) \) comes from Mann and Wald (1943). A random variable \( X \) is \( o_p(n^k) \) if \( n^{-k}X \to 0 \), and it is \( O_p(n^k) \) if \( n^{-k}X \) is bounded in probability.
where \( \theta^* \) will generally be different from (9) to (10), and also different from its previous usage, but still \( \theta^* \rightarrow 0 \) when \( \theta \rightarrow 0 \), as would be the case if the hypothesis is correct.

With \( I \) positive definite and \( F = F(\theta) \) having full column rank, \( P = (F'I^{-1}F) \) will be positive definite and can be factored with \( P^{1/2} \) as a symmetric \( p \times p \) positive definite matrix. Premultiplying through (7) by \( I^{-1/2} \) and through (8) by \( P^{-1/2} \), substituting (9) and (10) and setting \( \phi(\theta) = 0 \), will give after some rearrangement:

\[
I^{1/2}(\theta - \theta) = -I^{-1/2}[\partial^2 \phi/\partial \theta \partial \theta'(\theta^*)]I^{-1/2} - I^{-1/2}[F(\theta)]F^{-1/2}]^{-1} \left[ I^{-1/2} d(\theta) \right]
\]

\[
= \begin{bmatrix}
I_S & -M \\
-M' & 0
\end{bmatrix}^{-1} \begin{bmatrix}
I^{-1/2} d(\theta) \\
0
\end{bmatrix} + o_p(1)
\]

where \( M = I^{-1/2} F P^{-1/2} = I^{-1/2} F (F'I^{-1}F)^{-1} \). Note that \( M'M = I_p \) which implies that the elements of \( M \) are bounded. Using partitioned inversion,

\[
\begin{bmatrix}
I_S & -M \\
-M' & 0
\end{bmatrix}^{-1} = \begin{bmatrix}
I_S - MM' & -M \\
-M' & -I_p
\end{bmatrix}
\]

so that

\[
I^{1/2}(\theta - \theta) = (I_S - MM')I^{-1/2} d(\theta) + o_p(1)
\]

\[
P^{1/2} \lambda = -M'I^{-1/2} d(\theta) + o_p(1)
\]

\[
P \rightarrow N(0, I_p) .
\]

Expression (13) gives the asymptotic distribution, under the null hypothesis, of the Lagrange multipliers which arise in constrained maximum
likelihood estimation. One aspect of the relationship between constrained and unconstrained estimates which will be useful in the next section is given by subtracting (11) from (5) to give

$$I^{k_2} (\hat{\theta} - \theta) = MM'I^{k_2} d(\theta) + o_p(1). \tag{14}$$

1.3 The Testing Criteria

(i) The W Test

To assess the validity of the hypothesis $\phi(\theta) = 0$, the Wald test (W) considers whether $\phi(\hat{\theta})$ is significantly different from the zero vector [Wald (1943)]. When the hypothesized relationship is correct,

$$\phi(\hat{\theta}) = [F(\theta^*)]'(\hat{\theta} - \theta)$$

where $\theta^* \xrightarrow{P} \theta$, so that

$$P^{-k_2} \phi(\hat{\theta}) = (P^{-k_2}[F(\theta^*)]I^{-k_2})^{-1} I^{k_2} (\hat{\theta} - \theta)$$

$$= M'I^{k_2} (\hat{\theta} - \theta) + o_p(1)$$

$$\xrightarrow{d} N(0, I_p).$$

Then the quantity

$$[P^{-k_2} \phi(\hat{\theta})]'[P^{-k_2} \phi(\hat{\theta})] = [\phi(\hat{\theta})]'(P'I^{-1}F)^{-1}[\phi(\hat{\theta})] \tag{15}$$

will be asymptotically distributed as a $\chi^2$ random variable with $p$ degrees of freedom when the hypothesis is true, but will tend to have large values when $\phi(\theta) \neq 0$.

The quantity in (15) is not a statistic because $F = F(\theta)$ and $I = I(\theta)$ will generally depend upon the unknown parameters in $\theta$. One solution would be to replace unknown parameters with consistent estimates.
for which \( \hat{\theta} \) is the obvious choice. Then with \( P^{P-1}F_p + I_p \) and with \( P^{-1/2} \phi(\hat{\theta}) \) converging in distribution, the statistic

\[
W = [\phi(\hat{\theta})]'(\hat{F}'I^{-1/2}F_p)^{-1}[\phi(\hat{\theta})]
\]

(16)

will have an asymptotic distribution identical to that of (15). Using the limiting distribution to indicate approximate significance levels, the \( W \) test is performed by rejecting the null hypothesis if the calculated value of (16) exceeds the appropriate upper point of the \( \chi^2(p) \).

(ii) The LR Test

The likelihood ratio test (LR) compares the supremum of the likelihood when the hypothesized restrictions are imposed upon the parameters with that which is attainable when the restrictions are not enforced. Defining the ratio of constrained to unconstrained maximized likelihoods as \( \mu = L(\hat{\theta})/L(\theta) \), the LR statistic is

\[
LR = -2 \log \mu = 2[\ell(\hat{\theta}) - \ell(\theta)] .
\]

(17)

When the null hypothesis is correct, the asymptotic distribution of the LR statistic can be demonstrated from the expansion

\[
\ell(\hat{\theta}) = \ell(\theta) + \frac{1}{2}(\hat{\theta} - \theta)'[\theta^2 \partial / \partial \theta \theta'](\theta^*)'(\hat{\theta} - \theta)
\]

where \( \partial \ell / \partial \theta(\hat{\theta}) = d(\hat{\theta}) = 0 \) and \( \theta^* \xrightarrow{P} \theta \) when \( H_0 \) is true. Then from (17),

\[
LR = (\hat{\theta} - \theta)'I^{1/2}[I^{-1/2}[\theta^2 \partial / \partial \theta \theta'](\theta^*)]'I^{-1/2}I^{1/2}(\hat{\theta} - \theta)
\]

\[
= (\hat{\theta} - \theta)'I(\hat{\theta} - \theta) + o_p(1)
\]

from (4). Using (14) and noting that \( M'M = I_p \),

\[
LR = [d(\theta)]'I^{-1/2}MM'I^{-1/2}[d(\theta)] + o_p(1) .
\]

From the distributional statement in (13), it can be seen that the LR
statistic has an asymptotic $\chi^2(p)$ distribution when the null hypothesis is correct. The LR test is performed by rejecting the hypothesized restrictions for significantly large values of the statistic.

(iii) **The LM Test**

As its name implies, the Lagrange multiplier test (LM) examines the implicit costs or shadow prices of the imposed restrictions, as revealed by the calculated values of the multipliers. To test the hypothesis, the LM procedure considers whether the vector of Lagrange multipliers $\lambda$ is significantly different from the zero vector [Aitchison and Silvey (1958)]. From the null hypothesis distribution in (13), the quantity

$$[P^{1/2} \lambda]'[P^{1/2} \lambda] = \lambda'F'I^{-1}F\lambda$$

will be asymptotically $\chi^2(p)$ when $H_0$ is true, while large values would indicate that the sample evidence does not agree with the hypothesized relationship between the parameters.

For a practical test statistic, unknown parameters in $F$ and $I$ can be evaluated at the restricted estimates to give

$$LM = \lambda'F'I^{-1}F\lambda$$

(18)

where $\hat{F} = F(\hat{\theta})$ and $\hat{I} = I(\hat{\theta})$. Again the asymptotic distribution can be used to perform the test by rejecting the null hypothesis if the calculated value of the statistic exceeds the appropriate upper point of the $\chi^2(p)$ distribution.

---

4 Two other testing procedures which use criteria closely related to the LM statistic, the optimal $C(\alpha)$ tests of Neyman (1959) and the two-stage estimation approach of Durbin (1970), are considered in §3.3 and §4.2 respectively.
An alternative formulation of the LM statistic can be obtained by substituting for \( \tilde{T} \lambda \) in (18), from the first-order condition for restricted estimates given as (7) above. This form,

\[
LM = [d(\hat{\theta})]'I^{-1}[d(\hat{\theta})]
\]

(19)
is called the "efficient score statistic" after Rao (1948) and it is frequently the more convenient expression for use in practical applications. The LM criterion in this form can be considered as testing the significance of the difference from zero of the score vector when it is evaluated at the restricted parameter estimates.

Observe the interesting duality relationship between W and LM procedures. The W test uses unrestricted estimates with \( d(\hat{\theta}) = 0 \) and asks whether \( \phi(\hat{\theta}) \) is close to zero; the LM test uses restricted estimates with \( \phi(\hat{\theta}) = 0 \) and asks whether \( d(\hat{\theta}) \) is close to zero. It will be noted in §1.6 that a distinction between a test which requires \( \hat{\theta} \) and one which uses \( \hat{\theta} \) may be important if the unrestricted and restricted estimation problems have very different computational requirements.

In the W statistic (16) and the LM statistic (18) or (19), an estimate is required for the information matrix. Variants of these procedures are made possible by using matrices, other than \( I \) evaluated at the parameter estimates, in the quadratic forms which constitute the test criteria. Some \( s \times s \) nonsingular matrix \( H \), where

\[
I^{-1}_H P \rightarrow I_s
\]
could be used in place of \( I \) without affecting the asymptotic properties of the tests. Possible choices of \( H \) would include the (negative of the) Hessian matrix of second derivatives or its limiting form as \( n \rightarrow \infty \), evaluated at suitable consistent estimates of the unknown parameters if necessary.
1.4 Simplifications for Some Special Cases

So far, the hypothesized relationship between the parameters has been taken as \( \phi(\theta) = 0 \) for some, possibly nonlinear, function \( \phi(\cdot) \). A special case which arises frequently in practical problems is where the parameters are partitioned into two subsets as \( \theta' = (\theta_1', \theta_2') \), and the hypothesis asserts particular numerical values for the parameters in one subset as

\[
H_0: \quad \theta_1 = \theta_{10}
\]

or equivalently,

\[
H_0: \quad \phi(\theta) = [I_p : 0] \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} - \theta_{10} = 0.
\]

Only the \( p \) elements of \( \theta_1 \) enter the hypothesis and the \( (s-p) \) elements of \( \theta_2 \) are "nuisance parameters" in that their values are unknown under both null and alternative hypotheses. General parametric restrictions as \( \phi(\theta) = 0 \) can usually be reduced to (20) by a suitable reparameterization of the model, but this may be more complicated than direct application of the formulae in the previous section. However, for many practical situations the hypothesis of interest arises naturally in the form of (20), so it seems worthwhile to exploit whatever simplifications may be possible.

Define subvectors of the score vector and submatrices of the information matrix for a partitioning conformable with \( \theta' = (\theta_1', \theta_2') \):

\[
d(\theta) = \begin{bmatrix} d_1(\theta_1', \theta_2') \\ d_2(\theta_1', \theta_2') \end{bmatrix}
\]
Then the first-order conditions for an unrestricted maximum of the likelihood are

\[ \frac{d1(\hat{\theta}_1, \hat{\theta}_2)}{d2(V_1^1, V_2^1)} = 0 \]

and, from restricted estimation,

\[ \frac{d1(\theta_{10}, \theta_{20})}{d2(\hat{\theta}_1, \hat{\theta}_2)} = 0 \]

For the partitioned case, the three statistics defined in the previous section become:

\[ W = (\hat{\theta}_1 - \theta_{10})'(\hat{I}_{11})^{-1}(\hat{\theta}_1 - \theta_{10}) \]

\[ LR = 2[\lambda(\hat{\theta}_1, \hat{\theta}_2) - \lambda(\theta_{10}, \theta_{20})] \]

\[ LM = \lambda'\hat{I}_{11}\lambda = [d1(\theta_{10}, \theta_{20})]'\hat{I}_{11}[d1(\theta_{10}, \theta_{20})] \]

where \( \hat{I}_{11} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1} \) from partitioned inversion of \( I \), with \( \hat{I}_{11} \) and \( \hat{I}_{11} \) defined similarly.

One additional simplification is sometimes possible. If there are no information links between \( \theta_1 \) and \( \theta_2 \) so that \( I_{12} = I_{21} = 0 \), then \( I_{11} = I_{11}^{-1} \). In this case (where \( \hat{\theta}_1 \) is asymptotically distributed independently of \( \hat{\theta}_2 \)), construction of \( W \) and \( LM \) statistics is made easier because only part of the information matrix or its estimate needs to be obtained. Similarly, with general restrictions applying to a subset of the parameters that is related to the other parameters in the model by a
block-diagonal information matrix, only the relevant part of the information matrix is required.

1.5 Relative Power Properties

The three criteria defined in §1.3 all have the same limiting distribution in regular problems when the null hypothesis is correct. Of considerable interest, therefore, is the power of the testing procedures to discriminate against false null hypotheses.

Relative asymptotic power properties have been investigated by, inter alia, Wald (1943), Silvey (1959) and Moran (1970). All three principles, W, LR and LM will lead to consistent tests, i.e. for any fixed significance level, their power functions converge to unity as $n \to \infty$, for all $\phi(0) \neq 0$. Thus perfect discrimination is provided by any of the testing procedures with an infinitely large sample. But any reasonable test can be expected to have good power in large samples for alternative hypotheses far away from the null; it is only for alternatives where $\phi(0)$ is near the zero vector that the question of asymptotic power arises.

A more exacting basis for asymptotic comparisons is that of "local power", defined by Le Cam (1956). This considers the properties of the tests as $n \to \infty$, but with a sequence of alternative hypotheses which converges to the null in such a way that the statistics continue to have limiting distributions. Generalizing this idea to the framework of §1.2,

---

5 The notion of local asymptotic power for tests is analogous to that of asymptotic efficiency for estimators. With two root-n consistent estimators, $\hat{\theta}$ and $\tilde{\theta}$, comparing the limiting distributions of $\sqrt{n}(\hat{\theta} - \theta)$ and $\sqrt{n}(\tilde{\theta} - \theta)$ is more exacting than directly comparing distributions of $\hat{\theta}$ and $\tilde{\theta}$, which are both degenerate in the limit.
local alternatives can be specified as

$$H_A: \phi(\theta) = P^{-\frac{1}{2}} \gamma$$

where $P = (F'I^{-1}F)$ as before, and $\gamma$ is a $p$-vector of fixed elements.

The standard result from the analyses of asymptotic power cited above is that all three statistics have the same limiting distribution within the sequence of local alternatives, and that is the noncentral $\chi^2$ with $p$ degrees of freedom and noncentrality parameter

$$\gamma' \gamma = [P^{-\frac{1}{2}} \phi(\theta)]' [P^{-\frac{1}{2}} \phi(\theta)]$$

$$= [\phi(\theta)]' (F'I^{-1}F)^{-1} [\phi(\theta)].$$

Slightly different formulations of local alternatives may be easier to work with in some circumstances. For example, when $\lim n^{-1} I$ is assumed to be finite and with the partitioned situation of §1.4, the specification

$$H_A: \theta_1 = \theta_{10} + n^{-\frac{1}{2}} \gamma$$

leads to the noncentrality parameter of the asymptotic $\chi^2$ being

$$\lim_{n \to \infty} n^{-1} \gamma' [I_{11}]^{-1} \gamma = \lim_{n \to \infty} (\theta_1 - \theta_{10})' [I_{11}]^{-1} (\theta_1 - \theta_{10}).$$

Just as asymptotic efficiency is defined for estimators with limiting normal distributions relative to the Cramèr-Rao bound on the asymptotic covariance matrix, an optimality criterion for tests using statistics which are asymptotically $\chi^2$ has been defined by Neyman (1959) relative to the noncentrality parameter. Each of the three testing principles meets equally the criterion of asymptotically most powerful test.

By the asymptotic power criterion, there is nothing to choose between the three procedures for constructing tests, and very little is known about
their properties in finite samples. Most of the research in this direction has been confined to specific applications, with the usual conclusion being that relative power in finite samples depends upon the particular alternative hypothesis.

One general analysis of the problem, by Peers (1971), considers the simple null hypothesis $\theta = \theta_0$ against local alternatives $\theta = \theta_0 + o(n^{-1/2})$ for likelihoods formed from identically and independently distributed random variables. This analysis of relative power is not for the limiting case as $n \to \infty$, but includes terms in the power function to $o(n^{-1/2})$ by means of an Edgeworth expansion. Overall, the conclusion is that no one test is uniformly superior. In a direct comparison of $W$ and LM tests for $H_0: \theta = \theta_0$ where $\theta = (\theta_1, \theta_2)$ in a $N(\theta_1, \theta_2)$ model, it is shown that $W$ is more powerful whenever $\theta_2 < \theta_{20}$ but that LM is more powerful when $\theta_2 > \theta_{20}$. A bivariate example has observations jointly distributed as standard normal variables with unknown correlation coefficient, $\theta$. When $\theta > \theta_0 > 0$ and $\theta < \theta_0 < 0$, $W$ is shown to be more powerful than LM but the converse power relationship holds for other alternative hypotheses, except when $\theta_0 = 0$ in which case all three testing principles have the same properties to $o(n^{-1/2})$.

An interesting example of application of the three criteria which has been explored in considerable detail is that of testing whether frequencies of observations are the same over a number of classifications. For these tests of goodness of fit or tests of independence in a contingency table, the underlying probability model is the multinomial distribution. The familiar $\chi^2$ criterion,

$$\chi^2 = \sum_i (O_i - E_i)^2 / E_i$$
where \( O_i \) is the observed number and \( E_i \) is the expected number of observations in the \( i \)'th cell, will be given by the LM principle. [See Aitchison and Silvey (1960, p.167) or Rao (1973, p.442).]

A large number of studies have compared W, LR and LM tests in this situation, including usage of exact distributions as well as the asymptotic \( \chi^2 \) to set significance levels. Chapman (1976) gives references to many comparisons that have been made, but general conclusions are rare except that the asymptotic \( \chi^2 \) appears to be a more reliable guide to exact probabilities for LR than for the LM statistic in many cases. On the question of power, there seems to be some suggestion that the LR procedure might generally be better with simple alternative hypotheses, but when the alternative does not specify all of the parameters the properties of the three tests are quite comparable.

Lee (1971) examined the relative powers of W, LR and LM statistics in the forms that are the usual alternative criteria for testing the multivariate linear hypothesis. Cases were found in which W is more powerful than LR which is more powerful than LM, but equally there are situations for which the reverse ranking holds.

In summary, none of the tests is generally more powerful, even in the simple models for which exact results are available or for which higher order (than asymptotic) behaviour has been examined.\(^6\) Sometimes, with specific knowledge of the alternative hypothesis in a simple statistical model, it may be possible to prefer one testing procedure over the others for its better power properties in finite samples. But for econometrics, where alternatives are typically vaguely formulated

\(^6\) A familiar econometric situation in which all three criteria would give exactly the same decision when referred to their exact distributions (but not with asymptotic ones) is noted in §2.4.
and where statistical models tend to be quite complicated, there is little evidence in the literature to indicate superiority of any one of the three asymptotically equivalent testing principles.

1.6 Relative Computational Requirements

A feature of the three testing methods which serves to distinguish them is the different computational requirements for the tests to be performed. The $W$ test needs only the unrestricted estimation problem to be solved, the $LM$ test is based on the results of restricted estimation alone, whereas the LR procedure requires both sets of estimates.

Sometimes unrestricted estimation will be straightforward relative to the restricted situation. For example, to test a nonlinear relationship between the coefficients of a linear regression model, unrestricted maximum likelihood estimates will be given by ordinary least squares (OLS). Then the $W$ statistic can be constructed readily from the OLS results while restricted estimates for LR and LM tests will require iterative numerical methods to solve nonlinear estimating equations.

Another situation where the $W$ test would be most useful from the point of view of computational ease occurs when there are a number of possible sets of restrictions to be tested. Expanding on the results of Anderson (1971) for the normal linear model, Mizon (1977) argues the case for approaching a nested set of hypotheses by proceeding from the least restrictive specification and testing the hypotheses sequentially in increasing order of restrictiveness. With the least restrictive formulation as the maintained hypothesis, the $W$ test procedure would be the most appealing computationally as only the maintained general model need be estimated. The sequence of tests can then be performed by checking whether the unrestricted estimates satisfy the additional restrictions of
each null hypothesis, taken in turn.

In many situations, however, imposing the null hypothesis upon the parameters simplifies the model and its estimation. Econometric modelling typically proceeds by starting with a relatively simple statistical model and then performing diagnostic checks to see whether more complicated additions to the model should be considered. In this case, restricted estimation will be comparatively easy and, furthermore, restricted estimates will generally be available at the stage when such diagnostic checks are to be performed, so the LM procedure would be most attractive. Chapters 3 onward deal in some detail with applications of the LM test to diagnostic checking for various kinds of potential misspecification and to other situations where its relative economy of application would favour the LM method.

Both estimation problems have to be solved for the LR test to be applied. By the criterion of computational economy, this method is less attractive whatever the relative computational aspects of restricted and unrestricted estimation. However, if the hypothesis were to be rejected then presumably unrestricted estimates will be required anyway, whereas if it is accepted considerations of estimation efficiency would indicate that it be imposed. Thus if a W test were to accept an hypothesis or an LM test to reject it, then the computation involved would be the same as for the LR test. But when a number of tests are to be performed, as with an ordered sequence of hypotheses or for routine diagnostic checking, W and LM procedures respectively will generally be preferred for their ease of application.
1.7 Problems with Parameters on Boundaries

Standard asymptotic theory for maximum likelihood requires the true parameter value to be an interior point of some open subset of the parameter space. This, together with continuity conditions on the log-likelihood and its derivatives, permits the usual expansions by which the asymptotic distributions of maximum likelihood estimators are established from their relationship with the score. However, many interesting hypotheses assert that some of the parameters are on a boundary of the parameter space, and for these situations the standard maximum likelihood theory requires some modifications.

For simplicity, consider a one-parameter problem in which the parameter space is $\theta = \{\theta | 0 \leq \theta < \infty\}$ and for which root-n norming is appropriate for the maximum likelihood estimator to have a limiting distribution. If the null hypothesis is $\theta = \theta_0$, then the unrestricted estimate $\hat{\theta}$ is given by maximizing the likelihood over $\theta \in \Theta$. The $W$ test in this case would be based upon $(\hat{\theta} - \theta_0)$, but quite obviously $\sqrt{n}(\hat{\theta} - \theta_0)$ cannot be asymptotically normal, even when the hypothesis is correct.

Chernoff (1954) examines the effect upon the LR statistic when it is used to test hypotheses that parameters lie upon boundaries. For the one-parameter case, it is shown that LR is distributed under the null hypothesis as a random variable which is zero with probability $\frac{1}{2}$ and behaves like a $\chi^2(1)$ also with probability $\frac{1}{2}$. The only modification in this situation, for $W$ and LR critical points to be taken from the $\chi^2(1)$, is that the upper 2$\alpha$ percentage point of the $\chi^2$ distribution would be required for the test to have size $\alpha$ in large samples. [See also Cox and Hinkley (1974, pp.320-321).] Appropriate treatments for the LR statistic when the hypothesis sets more than one parameter to an edge
of the parameter space are also given by Chernoff (1954).

Interestingly enough, the LM test which is based upon the first derivative of the log-likelihood has been shown by Moran (1971) to be unaffected by the boundary value hypothesis situation. When the null hypothesis specifies more than one parameter, even if only one parameter is on the edge of the parameter space, Chant (1974) illustrates how maximum likelihood estimators can have nonnormal asymptotic distributions which are more complicated than simple truncation. But Moran's result on the LM test for the one-parameter case is shown to be applicable. Thus, while W and LR require modification for use in these nonstandard situations, the LM test has the usual properties.7

Hypotheses which set parameters onto boundaries are of interest in many econometric applications. Usually such parameters have interpretations as variances, e.g. in error components, random coefficient and other heteroscedastic generalizations of the regression model where the hypothesis of absence of one of these effects is expressed parametrically by setting a variance to zero. If variance estimates are constrained to be nonnegative by the estimation procedure, tests using W and LR criteria will require some modification but the LM statistic will have the same asymptotic properties as in the usual situation of an interior parameter point.

---

7 Moran (1971) and Chant (1974) refer to C(α) tests rather than the LM test. The former is described in §3.3 where its relationship to the LM procedure is discussed; it is sufficient for the present purpose to note that the LM test is a member of the class of optimal C(α) tests.
1.8 **Singular Information Matrices**

In the discussion so far, it has been assumed that the information matrix $I$ is positive definite. Sometimes however, the natural formulation of the unrestricted model gives a singular information matrix, either generally for all values of the parameters or for some regions of the parameter space that are of particular interest. If the information matrix is singular, the parameters are said to be unidentifiable. Obviously, the expressions that have been given for the test statistics which include $I^{-1}$ will require some modification in this situation.

Silvey (1959) and Aitchison and Silvey (1960) treat in considerable detail a method which enables the W and LM procedures to be applied when some of the restrictions in the hypothesis to be tested can be used to identify the parameters. Suppose that the information matrix for the $s$ parameters in $\theta$ has rank $(s-q)$ but that $q$ of the $p$ constraints in $\phi(\theta) = 0$ are sufficient for identifiability. Without loss of generality, the identifying restrictions can be taken to be the first $q$ members of $\phi(\theta)$. The Aitchison-Silvey approach is then to partition $F = \partial \phi / \partial \theta$ as $F = (F_1 : F_2)$ where $F_1$ is $s \times q$ and $F_2$ is $s \times (p-q)$, to give $(I + F_1'F_1)^{-1}$ as a nonsingular matrix. An appropriate modification to the W and LM tests is shown to be given by replacing $I^{-1}$ where it appears in the formulae for the statistics by $(I + F_1'F_1)^{-1}$, and making reference to the $\chi^2(p-q)$ distribution rather than the $\chi^2(p)$ for significance probabilities. "Unrestricted" estimates for use in the modified W test are obtained in Aitchison and Silvey (1960, p.163) by imposing the $q$ identifying restrictions.

---

8 Examples are noted in §3.7 and §4.4.
Because the LM test requires only those estimates which are obtained with all of the constraints imposed, modifying it to allow for singularity of the information matrix is usually quite straightforward. However there is an interpretation of the modified LM statistic that will make its usage even simpler in many situations. Note that \((I + F'F)^{-1}\) is a generalized inverse of \(I\), i.e. a g-inverse in the basic sense, defined for matrix \(A\) as \(A^{-}\) where \(AA^{-}A = A\). This can be seen by observing that when \(I\) is positive semidefinite of rank \((s-q)\), it can be written as \(I = C'C\) where \(C\) is an \((s-q) \times s\) matrix of full row rank; that \((I + F'F)^{-1}\) is a g-inverse of \(I\) follows directly from Rao and Mitra (1971, Complement 5(b), p.40). The Aitchison-Silvey modified LM statistic for singular information matrices is, therefore,

\[
LM = [d(\tilde{\theta})]' \tilde{I}^{-} [d(\tilde{\theta})]
\]

(21)

where \(\tilde{I}^{-}\) is some g-inverse of \(\tilde{I}\).

Generalized inverses by the basic definition are not unique, and this raises the question of the effect upon (21) of different choices of g-inverse. From Rao and Mitra (1971, Lemma 2.2.4(ii), p.21), a quadratic form as in (21) is invariant to the choice of g-inverse if and only if \(d(\tilde{\theta})\) is contained in the column space of \(\tilde{I}\), i.e. iff there exists an \(s\)-vector, \(f\), such that \(d(\tilde{\theta}) = \tilde{I}f\). When \(d(\theta)\) is normally distributed, or in large samples when it is approximately normally distributed, with mean vector zero and covariance matrix \(I\) of rank \((s-q)\), it can be represented as a singular linear transformation from a smaller dimensioned vector of random variables with a nonsingular covariance matrix, e.g.

\[
d(\theta) = C'x \text{ where } x \sim N(0, I_{s-q}).
\]

Then \(d(\theta)\) is (asymptotically) contained within the column space of \(C'\) and hence that of \(I = C'C\), implying that there exists some vector \(f\) for which \(d(\tilde{\theta}) = \tilde{I}f\).
Whether the modified LM statistic (21) is numerically invariant to the choice of g-inverse will depend upon the structure of the problem and, in particular, upon what asymptotic approximations may be involved in estimating the information matrix for use in the quadratic form which constitutes the statistic. It is clear that asymptotic properties are unaffected by the choice of g-inverse; in the two applications using the modified LM statistic in this thesis, exact invariance holds.

1.9 Two-Step Estimators

Much recent research in econometrics has been directed toward devising estimators which have asymptotic distributions coinciding with that of the maximum likelihood estimator, but which can be obtained in a fixed number of well-defined operations. The idea behind the so-called efficient two-step estimation methods is that of linearizing nonlinear normal equations for a maximum of the likelihood, using initial estimates from a preliminary step. These first step estimates are required to be consistent but not necessarily efficient. Efficient estimates are given from the second step, which usually coincides with one round of a numerical estimation algorithm for maximizing the likelihood, starting with the consistent estimates as the initial trial solution.

Since several references are made in the sequel to nonlinear estimation algorithms and to two-step estimators, the general procedures will be sketched briefly here. For this discussion, the more familiar root-n norming is assumed to be appropriate. Thus it is implicit that

\[ n^{-1/2} d(0) \xrightarrow{D} N(0, \lim_{n \to \infty} n^{-1} I) \]

to give the distribution of the unrestricted maximum likelihood estimator.
Different situations will suggest different approaches to the first step, which is to obtain root-\(n\) consistent estimates, so details of this step are not considered. Denote by \(\hat{\theta}\) the vector of estimates which are consistent for the parameters of the unrestricted model, whether the hypothesis is true or not. A class of efficient two-step estimators can be defined by

\[
\hat{\theta} = \bar{\theta} + [H(\bar{\theta})]^{-1} d(\bar{\theta})
\]

(22)

where \(H(\theta)\) is an \(s \times s\) nonsingular matrix such that \(I^{-1}H + I_s\). Then a proof along the lines of Rothenberg and Leenders (1964, p.69) can generally be used to show that

\[
n^\frac{1}{2}(\hat{\theta} - \theta) = n^\frac{1}{2}(\hat{\theta} - \theta) + o_p(1)
\]

(23)

implying that the two-step estimator has the same limiting distribution as the maximum likelihood one.

For hypothesis testing, similar methods could also be applied to the constrained estimation problem when the model is restricted by the null hypothesis, to give two-step estimators which are consistent and efficient when the hypothesis is correct and for local deviations from it. These estimates will be denoted by \(\tilde{\theta}\) and will be related to the constrained maximum likelihood estimate \(\bar{\theta}\) by

\[
n^\frac{1}{2}(\tilde{\theta} - \theta) = n^\frac{1}{2}(\hat{\theta} - \theta) + o_p(1)
\]

(24)

---

9 An estimator \(\bar{\theta}\) is root-\(n\) consistent if \(\sqrt{n}(\bar{\theta} - \theta) = o_p(1)\), implying \(\bar{\theta} \overset{p}{\to} \theta\).
under the null hypothesis and local alternatives.

Various two-step estimation procedures can be distinguished by their choice of matrix $H(\theta)$ and thereby linked to different numerical estimation algorithms. With $H(\theta) = I(\theta)$, the formula in (22) represents one round of Fisher's scoring algorithm; for example, Harvey (1976) proposed an efficient two-step estimator for a heteroscedastic regression model, based upon one round of scoring. An alternative procedure could have $H(\theta) = [-\partial^2 \ell / \partial \theta \partial \theta]'(\theta)$, so that (22) corresponds to a round of Newton-Raphson; this choice was suggested by Rothenberg and Leenders (1964) as a relatively simple way of obtaining efficient estimates of the parameters of a simultaneous equations system. Frequently, the log-likelihood is essentially a sum of squares:

$$\ell(\theta) = \sum_{t=1}^{n} f_t^2(\theta) + k_2$$

where $k_1$ and $k_2$ are constants and $f_t(\theta)$ is a scalar function of $\theta$. Hartley and Booker (1965) show that, under suitable conditions, the matrix

$$H(\theta) = \sum_{t=1}^{n} \partial f_t(\theta) / \partial \theta (\partial f_t(\theta) / \partial \theta)'$$

as in the Gauss-Newton algorithm, can be used to obtain efficient two-step estimates from (22). As an example of econometric application, the two-step estimator devised by Hatanaka (1974) for the dynamic regression model with autoregressive errors can be obtained in this way.

From (23), it is clear that the asymptotic properties of the $W$ test would be unaffected by use of efficient two-step estimates in place of maximum likelihood ones in the formula for the statistic. It is not difficult to show that $d(\tilde{\theta}) = d(\tilde{\theta}) + o_p(1)$ when (24) holds, so that the
LM test based upon efficient estimates of the restricted model, other than constrained maximum likelihood, would continue to have the same limiting properties. Furthermore, from (23) and (24) it can be shown that $\ell(\hat{\theta}) = \ell(\hat{\theta}) + o_p(1)$ and $\ell(\tilde{\theta}) = \ell(\tilde{\theta}) + o_p(1)$ under the null hypothesis and for local alternatives, implying that the LR formula using efficient estimates other than maximum likelihood has the same asymptotic properties.
CHAPTER 2
TESTING COEFFICIENT CONSTRAINTS IN LINEAR REGRESSION*

2.1 Introduction

The linear regression relationship, including possibly a nonscalar covariance matrix for the disturbances, is the most commonly used statistical model in econometrics. In this chapter, the inference principles introduced in Chapter 1 are applied to the problem of testing hypotheses about the coefficient parameters in a regression model. This development is intended to serve two purposes: one is to exposit the procedures in a familiar econometric setting, the other is to explore further relationships among the three asymptotically equivalent criteria in models which have wide use in econometric practice.

If, in some general problem, the three principles $W$, $LR$ and $LM$ were to be used for testing of the same hypothesis, then the same critical value from the asymptotic $\chi^2$ distribution would be appropriate for all three criteria. Given a finite amount of data from which a model is estimated and the statistics are computed, however, the numerical values obtained for the three statistics will generally be different. Then there will be some significance levels at which the asymptotically equivalent testing procedures give conflicting inferences. For certain hypotheses within two particular forms of the general linear model, Savin (1976) and Berndt and Savin (1977) show that a systematic numerical inequality will pertain among the computed values of the three statistics, viz.

$$W > LR > LM.$$  \hspace{1cm} (1)

This inequality, its extension to other situations and its implications

* This chapter is based in part upon material to be published as Breusch (1979).
are the main concerns of this chapter.

In §2.2 the three statistics are derived for hypotheses in the form of linear constraints upon the coefficients of the general linear model with nonscalar disturbance covariance matrix. An interesting interpretation of the W and LM procedures is noted in §2.3, allowing the inequality (1) to be deduced immediately for a class of models which includes both the Savin and the Berndt and Savin specifications as special cases. Application to the linear model with scalar covariance matrix is considered in §2.4; in this situation, setting significance levels by exact rather than asymptotic distributions leads to the same decision, whichever of the three testing principles is used. Possible extensions of the inequality relationship are entertained in §2.5 to §2.7. Nested sequences of hypotheses are considered in §2.5; nonlinearities either in the unrestricted regression model or in the hypothesized relationship between the parameters are examined in §2.6, and in §2.7 situations are treated in which the coefficient parameters are related to those in the disturbance covariance matrix. Finally, §2.8 contains a discussion of the significance of the inequality relationship.

2.2 Notation and Derivation of the Statistics

Consider the linear regression model

\[ y = X\beta + \epsilon \]  

(2)

where \( y \) is an \( n \)-vector of observations on the dependent variable, \( X \) is an \( n \times k \) matrix of observations on explanatory variables, \( \beta \) is a \( k \)-vector of unknown coefficient parameters, and the \( n \)-vector of disturbances, \( \epsilon \), is distributed as

\[ \epsilon \sim N(0, \Omega) \]  

(3)

The elements of the covariance matrix \( \Omega \) are assumed to be functions of
a fixed number of unknown parameters, say $\Omega = \Omega(\alpha)$, where there are no functional relationships between the $\alpha$ parameters and the $\beta$ regression coefficients. It is assumed that $\Omega$ is positive definite, that its parameterization permits $\alpha$ parameters to be identified, and that rank($X$) = $k < n$ so that $\beta$ parameters are also identified. Regressors are either exogenous or, if lags of the dependent variable are included in $X$, the structure of the covariance matrix is such that the off-diagonal blocks of the information matrix connecting $\beta$ and $\alpha$ are composed entirely of zeros, i.e. $I_{\beta\alpha} = I_{\alpha\beta}' = 0$. This implies that the maximum likelihood estimate of $\beta$ will be asymptotically distributed independently of that of $\alpha$. Apart from the foregoing conditions, the precise nature of the parameterization of $\Omega$ is not presently of interest so the unknown parameters in the model will be referred to as $\beta$ and $\Omega$.

The hypothesis to be tested specifies $p$ independent linear restrictions on $\beta$, written as

$$H_0: F'\beta = f$$

where $F$ is a $k \times p$ matrix with rank($F$) = $p < k$.

For the unrestricted model, the log-likelihood is given by

$$\ell(\beta, \Omega) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \log |\Omega| - \frac{1}{2} (y - X\beta)' \Omega^{-1} (y - X\beta).$$

(5)

Full details of estimation will depend upon the parameterization of the covariance matrix, but are not required for the present purpose. With the information matrix block diagonal between $\beta$ and the parameters in $\Omega$, and with the hypothesis referring to $\beta$ alone, only estimation of $\beta$ is of particular interest. The components of the score vector and information

---

1 This assumption would exclude from present consideration most models in which a nonscalar covariance matrix implies autocorrelated disturbances, when lags of the dependent variable are included among the regressors.
matrix required for testing $H_0$ are then
\[ d_\beta(\beta, \Omega) = X'\Omega^{-1}(y-X\beta) \]
\[ \lambda_\beta = X'\Omega^{-1}z . \]

Unrestricted estimates which jointly maximize (5) when the hypothesis is ignored will be denoted by $(\hat{\beta}_u, \hat{\Omega})$ where, from setting (6) to zero,
\[ \hat{\beta}_u = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y . \]

Restricted estimates which jointly maximize the log-likelihood subject to the constraints on $\beta$ imposed by the hypothesis will be denoted by $(\tilde{\beta}_R, \tilde{\Omega})$. Appending $\lambda'(F'\beta-f)$ to (5) gives the Lagrangian function,
\[ \psi(\beta, \Omega, \lambda) = \lambda(\beta, \Omega) + \lambda'(F'\beta-f) \]
for which the first-order conditions on $\beta$ and $\lambda$ for a maximum at $(\tilde{\beta}_R, \tilde{\Omega}, \tilde{\lambda})$ are
\[ X'\tilde{\Omega}^{-1}(y-X\tilde{\beta}_R) + F\tilde{\lambda} = 0 \]
\[ F'\tilde{\beta}_R - f = 0 . \]

Solving these for $\tilde{\beta}_R$ gives
\[ \tilde{\beta}_R = \hat{\beta}_R + (X'\Omega^{-1}X)^{-1}F\tilde{\lambda} \]
(7)
where
\[ \hat{\beta}_R = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y \]
(8)
\[ \tilde{\lambda} = -A^{-1}(F'\hat{\beta}_R-f) \]
\[ A = F'(X'\Omega^{-1}X)^{-1}F = F'I_{\beta\beta}^{-1}F \]
with $\hat{\Omega}$ the same as $\Omega$ but with $\hat{\Omega}$ replacing $\Omega$. Note that the hat indicates an unrestricted estimate while the tilde
indicates an estimate from imposing the hypothesized constraints upon the $\beta$ coefficients. However the construction in (8) would be the unrestricted maximum likelihood estimate of $\beta$ if $\hat{\Omega}$ was the known value of the covariance matrix, i.e. it is the generalized least squares (GLS) estimate with $\hat{\Omega}$ as the covariance matrix. Subscripts on an estimate of $\beta$ indicate that it can be interpreted as the maximum likelihood estimate (or GLS estimate) conditional upon $\hat{\Omega}$ (subscript u) or $\tilde{\Omega}$ (subscript R). This gives $\hat{\beta}_u$ and $\hat{\beta}_R$ as the notation for the proper unrestricted and restricted maximum likelihood estimates respectively, when $\Omega$ is estimated jointly with $\beta$ in each case. Another construction which will prove useful is

$$\tilde{\beta}_u = \hat{\beta}_u - (X'\hat{\Omega}^{-1}X)^{-1}F\hat{\beta}_u - f$$

(9)

where $\hat{\Omega}$ is $\Omega$ with $\hat{\Omega}$ replacing $\Omega$. The interpretation of $\tilde{\beta}_u$, which satisfies $F'\tilde{\beta}_u = f$, is that it is the constrained estimate of $\beta$, conditional upon $\hat{\Omega}$ being the known covariance matrix.

Residuals from the proper unrestricted and restricted estimates are denoted by

$$\hat{\epsilon}_u = y - X\hat{\beta}_u \quad \text{and} \quad \hat{\epsilon}_R = y - X\hat{\beta}_R .$$

Similarly, residuals from the constructed "estimates" can be defined as

$$\tilde{\epsilon}_R = y - X\tilde{\beta}_R \quad \text{and} \quad \tilde{\epsilon}_u = y - X\tilde{\beta}_u .$$

The three statistics for testing the hypothesis $F'\beta = f$ can now be given directly from the formulae in §1.3:

(i) \[ W = (F'\hat{\beta}_u - f)'(F'F)^{-1}(F'\hat{\beta}_u - f) \]

\[ = (F'\hat{\beta}_u - f)'A^{-1}(F'\hat{\beta}_u - f) . \]
But, from (9),

\[ \hat{\varepsilon}_u = y - X\hat{\beta}_u \]

\[ = \hat{\varepsilon}_u + X(X'\hat{\Omega}^{-1}X)^{-1}F\hat{\alpha}^{-1}(F'\hat{\beta}_u - f) \]

and noting that \( X'\hat{\Omega}^{-1}\hat{\varepsilon}_u = 0 \),

\[ \hat{\varepsilon}_u = \hat{\varepsilon}_u + (F'\hat{\beta}_u - f)'\hat{\alpha}^{-1}(F'\hat{\beta}_u - f) \].

Alternatively,

\[ W = \hat{\varepsilon}'_u \hat{\Omega}^{-1}\hat{\varepsilon}_u - \hat{\varepsilon}'_u \hat{\Omega}^{-1}\hat{\varepsilon}_u \]

which is another form of the \( W \) statistic.

(ii) \( LR = 2[\lambda(\hat{\beta}_u, \hat{\Omega}) - \lambda(\hat{\beta}_R, \tilde{\Omega})] \)

\[ = \log |\hat{\Omega}| + \hat{\varepsilon}'_R \hat{\Omega}^{-1}\hat{\varepsilon}_R - \log |\tilde{\Omega}| - \hat{\varepsilon}'_R \tilde{\Omega}^{-1}\hat{\varepsilon}_R . \]

(iii) \( LM = \lambda'F'\tilde{\Omega}^{-1}F\lambda \)

\[ = \lambda'\tilde{\Omega} \hat{\alpha} \].

Now, from (7),

\[ \tilde{\varepsilon}_R = y - X\tilde{\beta}_R \]

\[ = \hat{\varepsilon}_R - X(X'\hat{\Omega}^{-1}X)^{-1}F\tilde{\alpha} \]

and noting that \( X'\hat{\Omega}^{-1}\hat{\varepsilon}_R = 0 \),

\[ \tilde{\varepsilon}'_R \hat{\Omega}^{-1}\tilde{\varepsilon}_R = \hat{\varepsilon}'_R \tilde{\Omega}^{-1}\hat{\varepsilon}_R + \hat{\lambda}'\tilde{\Omega} \tilde{\alpha} \].

Alternatively,

\[ LM = \tilde{\varepsilon}'_R \tilde{\Omega}^{-1}\tilde{\varepsilon}_R - \tilde{\varepsilon}'_R \tilde{\Omega}^{-1}\tilde{\varepsilon}_R \]

which is another form of the \( LM \) statistic.
2.3 The Inequality Relationship

The inequality relationship (1) between the three criteria given in (10), (11) and (12) will follow from an interpretation of the \( W \) and \( LM \) procedures as conditional likelihood ratio tests and the relationship between conditional and unconditional optimization. With \( \hat{\Omega} \) set equal to the estimate \( \hat{\Omega} \) from full unrestricted maximum likelihood estimation, application of the likelihood ratio principle gives the \( W \) statistic. For notational brevity, the parameter space of \( \beta \) under the null hypothesis is denoted by \( B_o \equiv \{ \beta | F'\beta = f \} \). Then, from (10),

\[
W = \tilde{e}_u^{\hat{\beta} - 1} e_u^{\hat{\beta} - 1} - \tilde{e}_u^{\hat{\beta} - 1} e_u^{\hat{\beta} - 1} \\
= 2[\ell(\hat{\beta}_u, \hat{\Omega}) - \ell(\tilde{\beta}_u, \hat{\Omega})]
\]

where

\[
\ell(\hat{\beta}_u, \hat{\Omega}) = \sup_{\beta \in B_o} \ell(\beta | \hat{\Omega})
\]

\[
\leq \sup_{\Omega, \beta \in B_o} \ell(\beta, \Omega) = \ell(\hat{\beta}_R, \hat{\Omega})
\]

which implies \( W \geq LR \). Similarly, the \( LM \) statistic can be interpreted as arising from an application of the likelihood ratio principle, conditional upon \( \hat{\Omega} \) being set equal to the estimate \( \hat{\Omega} \) from restricted estimation under the null hypothesis. From (12),

\[
LM = \tilde{e}_R^{\hat{\beta}_R - 1} e_R^{\hat{\beta}_R - 1} - \tilde{e}_R^{\hat{\beta}_R - 1} e_R^{\hat{\beta}_R - 1} \\
= 2[\ell(\hat{\beta}_R, \hat{\Omega}) - \ell(\tilde{\beta}_R, \hat{\Omega})]
\]

where

\[
\ell(\hat{\beta}_R, \hat{\Omega}) = \sup_{\beta} \ell(\beta | \hat{\Omega})
\]

\[
\leq \sup_{\Omega, \beta} \ell(\beta, \Omega) = \ell(\hat{\beta}_u, \hat{\Omega})
\]
which implies \( LR \geq LM \).

Savin (1976) established the inequality relationship between the three criteria for testing linear constraints on \( \beta \) in the univariate regression model with exogenous regressors and first-order autoregressive disturbance:\(^2\)

\[
y_t = x_t' \beta + \varepsilon_t
\]
\[
\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad |\rho| < 1
\]

with \( u_t \sim NID(0, \sigma^2) \) for \( t = 1, \ldots, n \). In the present notation, this model fits the specification (2) and (3) where \( \Omega \) would be given by expression (6) of Savin (1976, p.1304).

Berndt and Savin (1977) considered the multivariate linear regression model

\[
y_t = B' z_t + \varepsilon_t
\]

where \( y_t \) is an \( M \)-vector of dependent variables, \( z_t \) is a \( K \)-vector of predetermined variables so that the \( B \) matrix of coefficients is \( K \times M \), with the \( M \)-vector of disturbances \( \varepsilon_t \sim NID(0, \Sigma) \) for \( t = 1, \ldots, T \).

All \( T \) observations can be written compactly as

\[
Y = ZB + E \tag{13}
\]

where \( y'_t, z'_t \) and \( \varepsilon'_t \) are the \( t \)'th rows of \( Y, Z \) and \( E \) respectively. An alternative representation is given by vectorizing both sides of (13) to obtain

\[
y = X\beta + \varepsilon
\]

\(^2\) Savin also examined this model with the first lag of the dependent variable included as an element of \( x_t \). The generalization of this situation is considered in §2.7 below.
with \( y = \text{vec} \, Y, \, X = (I_M \otimes Z), \, \beta = \text{vec} \, B \) and \( \varepsilon = \text{vec} \, E. \) This model also fits the specification of §2.2 with \( n = MT, \, k = MK \) and \( \Omega = (E \otimes I_T). \) Using \( \varepsilon' \Omega^{-1} \varepsilon = \text{tr}[\Sigma^{-1}E'E], \) the three statistics in this case can be written as

\[
W = T \text{tr} \left[ (\hat{E}'\hat{E})^{-1}(\hat{E}'\hat{E} - \hat{E}'\hat{E}) \right]
\]

\[
LR = T \log \left[ \left| \tilde{E}_R^T \tilde{E}_R \right| / \left| \hat{E}'\hat{E} \right| \right]
\]

\[
LM = T \text{tr} \left[ (\hat{E}_R^T \hat{E}_R)^{-1}(\hat{E}_R^T \hat{E}_R - \hat{E}'\hat{E}) \right]
\]

where \( \tilde{e}_u = \text{vec} \, \tilde{E}_u, \, \tilde{e}_R = \text{vec} \, \tilde{E}_R \) and \( \hat{e}_u = \hat{e}_R = \hat{e} = \text{vec} \, \hat{E}. \) No subscript is required on estimates and residuals from unrestricted estimation in this case, since the same estimates would be given by GLS (=OLS), whatever initial estimate of \( \Omega \) was used. These matrix forms for the statistics parallel closely the corresponding expressions in Berndt and Savin (1977).

The framework of §2.2 includes as special cases these two specifications in which the inequality had previously been established, but it is also considerably more general. Also included in the general framework would be the single equation model with other forms of nonspherical disturbances such as alternative autocorrelation patterns and most of the estimable heteroscedasticity formulations.\(^3\) The seemingly unrelated equations model, of which the multivariate regression can be considered to be the particular case of identical regressors in every equation, also fits (2) and (3) with \( \Omega = (E \otimes I_T) \) when the model is written in vector form, although \( X \) will not generally have a representation as a Kronecker product. As a further example of the greater generality of the result

---

\(^3\) Savin (1976) in his concluding comments states without proof the extension of the result to other nonscalar matrix representations in the univariate linear model.
established here, observe that (2) and (3) could be the vector form of a multivariate regression model (or a seemingly unrelated equations system) where "...the disturbances are no longer independently and identically distributed...", which was suggested as a "...topic for further research" by Berndt and Savin (1977, p.1276).

With a systematic numerical ranking existing between the values that would be calculated for the three asymptotically equivalent statistics, the possibility arises that the criteria will lead to conflicting inferences when referred to their common asymptotic critical value. This observation is considered in more detail in §2.8.

2.4 Scalar Covariance Matrix

Consider now the model of §2.2 where it is assumed that the disturbance covariance matrix is scalar, i.e. $\Omega = \sigma^2 I_n$. For the unrestricted model, maximum likelihood estimates are given by OLS and these will be denoted by $(\hat{\beta}, \hat{\sigma}^2)$ where $\hat{\sigma}^2 = n^{-1} \hat{\varepsilon}'\hat{\varepsilon}$ with $\hat{\varepsilon} = y - X\hat{\beta}$ as the OLS residuals. Restricted least squares estimation is well documented, e.g. in Theil (1971, Sects 1.8 and 3.7), so it will suffice to define the notation $(\hat{\beta}, \hat{\sigma}^2)$ and to observe that $\hat{\sigma}^2 = n^{-1} \hat{\varepsilon}'\hat{\varepsilon}$ where $\hat{\varepsilon} = y - X\hat{\beta}$. Note that subscripts on estimates can be suppressed because estimation of $\beta$ does not require an estimate of the covariance matrix.

To test linear restrictions as in (4), the three criteria in this case would be, from (10), (11) and (12) respectively:

$$W = (\hat{\varepsilon}'\hat{\varepsilon} - \hat{\varepsilon}'\hat{\varepsilon})/\hat{\sigma}^2$$

$$= n(\hat{\varepsilon}'\hat{\varepsilon} - \hat{\varepsilon}'\hat{\varepsilon})/\hat{\varepsilon}'\hat{\varepsilon}$$

$$LR = n(\log \hat{\sigma}^2 - \log \hat{\sigma}^2)$$

$$= n \log (\hat{\varepsilon}'\hat{\varepsilon}/\hat{\varepsilon}'\hat{\varepsilon})$$
\[ LM = (\hat{c}'\hat{c} - \hat{c}'\hat{c})/\hat{\sigma}^2 \]

\[ = n(\hat{c}'\hat{c} - \hat{c}'\hat{c})/\hat{c}'\hat{c}. \]

Taking the W statistic as an arbitrary reference point, the other two criteria can be expressed as transformations to it:

\[ W = n[\exp(n^{-1}LR) - 1] \]

\[ W = n LM/(n - LM). \]

Both of these transformations are one-to-one, indicating that if exact distributions rather than asymptotic ones were to be employed to judge the significance of calculated values of the three criteria, then exactly the same decision would be reached in each case. With a scalar covariance matrix (corresponding to \( M = 1 \) in (13)), there will be no conflict among the criteria when exact distributions are used.

If there are lagged values of the dependent variable included among the explanatory variables, the exact distribution will generally be quite intractible. However, when the regressors can all be taken as fixed, use of exact distributions would lead to a test based upon

\[ F(p, n-k) = \frac{(\hat{c}'\hat{c} - \hat{c}'\hat{c})/p}{\hat{c}'\hat{c}/(n-k)} = \frac{(n-k)}{np} W \]

being distributed as \( F \) with \( p \) and \( (n-k) \) degrees of freedom, where \( p \) is the number of effective independent restrictions and \( (n-k) \) is the degrees of freedom with which the variance is estimated in the unrestricted model. Thus all three principles \( W, LR \) and \( LM \), when used with exact significance points, would lead to the familiar F-test which is uniformly most powerful within wide classes of inference procedures. [See, \textit{e.g.}, Seber (1966, Ch. 4).]
2.5 Nested Hypotheses

For the multivariate linear regression model, Berndt and Savin (1977) discuss the nested hypothesis situation where one set of independent linear restrictions is subsumed into the model and it is desired that an additional independent set of restrictions be tested. It will now be shown that the inequality relation between the criteria continues to hold in this case, for the more general class of specifications of §2.2.

Suppose that the model with \( k \) regressors is

\[
y = X\beta + \epsilon
\]

(14)

where the disturbance covariance matrix may be nonscalar as in §2.2. The \( p_1 \) restrictions in

\[
F_1'\beta = f_1
\]

are to be maintained, and an additional set of \( p_2 \) restrictions in

\[
H_0 : F_2'\beta = f_2
\]

(15)

are to be tested. By the independence assumption between restrictions, \( \text{rank}(F_1) = p_1, \text{rank}(F_2) = p_2 \) and \( \text{rank}(F_1 : F_2) = (p_1 + p_2) \).

The model with the maintained restrictions incorporated can be reparameterized to be of the same form as (2) above with new dependent variable, regressors and coefficients. Partition \( F_1 \) into the first \( p_1 \) and other \((k-p_1)\) rows, and partition \( \beta \) and \( X \) conformably, i.e.

\[
F_1 = \begin{bmatrix} F_{11} \\ F_{12} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad X = (X_1 : X_2)
\]

where \( F_{11} \) can be taken to be nonsingular by suitable reordering of the parameters in \( \beta \) if necessary. Maintained restrictions \( F_1'\beta = f_1 \) can
then be written as freedom equations:

\[ \beta = \begin{bmatrix} (F'_{11})^{-1}(f_1 - F'_{12}\beta_2) \\ \beta_2 \end{bmatrix} . \] (16)

Substituting (16) into the unrestricted model (14) gives the maintained model as

\[ y = (X_1 : X_2) \begin{bmatrix} (F'_{11})^{-1}(f_1 - F'_{12}\beta_2) \\ \beta_2 \end{bmatrix} + \varepsilon . \]

Alternatively,

\[ [y - X_1(F'_{11})^{-1}f_1] = [X_2 - X_1(F'_{11})^{-1}F'_{12}]\beta_2 + \varepsilon \]

\[ y^* = X^*\beta_2 + \varepsilon \] (17)

which is exactly the same form as the specification (2) of §2.2.

In a similar way, the additional set of restrictions in (15) which are to be tested can be rearranged to apply linearly to the \( \beta_2 \) coefficients of (17). Partition \( F_2 \) into the first \( p_1 \) and other \( (k-p_1) \) rows so that the hypothesis becomes

\[ (F'_{21} : F'_{22}) \begin{bmatrix} (F'_{11})^{-1}(f_1 - F'_{12}\beta_2) \\ \beta_2 \end{bmatrix} = f_2 . \]

Alternatively,

\[ [F'_{22} - F'_{21}(F'_{11})^{-1}F'_{12}]\beta_2 = f_2 - F'_{21}(F'_{11})^{-1}f_1 \]

\[ (F^*)'\beta_2 = f^* \]

which has exactly the same form as (4).

At each testing stage through a nested sequence of linear hypotheses,
the maintained model and the hypothesis under test can be arranged into
the general format of §2.2. Provided the combined regressors in \( X^* \)
and the assumed form of the covariance matrix give a block-diagonal
information matrix, the computed values of \( W, LR \) and \( LM \) will be
related by the inequality for each test.

2.6 Nonlinearities

It is of some interest to consider the possibility of the inequality
relationship among the three testing criteria extending to the case
where the \( \beta \) parameters enter the model nonlinearly and/or the hypothesis
to be tested specifies nonlinear relationships among the elements of \( \beta \).

Firstly consider the linear regression model with non-scalar covariance matrix as in §2.2 but suppose the hypothesis is

\[
H_0: \phi(\beta) = 0
\]

where the \( p \) constraints in \( \phi(\beta) \) include some nonlinear relationships.
Then the \( k \times p \) matrix \( F(\beta) = \partial \phi(\beta)/\partial \beta \) will be a function of the \( \beta \)
parameters. The first-order conditions on \( \beta \) and the Lagrange multipliers
in \( \lambda \) for a restricted maximum of the likelihood are then

\[
X'\hat{\Omega}^{-1}(y - X\hat{\beta}_R) + \hat{F}\lambda = 0
\]

\[
\phi(\hat{\beta}_R) = 0.
\]

Solving for \( \hat{\beta}_R \) will give

\[
\hat{\beta}_R = \hat{\beta}_R + (X'\hat{\Omega}^{-1}X)^{-1}\hat{F}\lambda
\]

where, as before,

\footnote{These possibilities are speculated upon by Berndt and Savin (1977, p.1276) and Mizon (1977, p.1229).}
\[ \hat{\beta}_R = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\eta}^{-1}y \]

but now, with nonlinear restrictions, the multipliers will not be linear in \( \hat{\beta}_R \), and \( \tilde{F} = F(\hat{\beta}_R) \) will depend upon estimates.

The LM test criterion can be formulated as

\[ LM = \lambda' \tilde{F}'(X'\hat{\Omega}^{-1}X)^{-1}\tilde{F} \lambda \]

\[ = \tilde{\varepsilon}' \hat{\Omega}^{-1} \tilde{\varepsilon} - \hat{\varepsilon}' \hat{\Omega}^{-1} \hat{\varepsilon} \]

\[ = 2[\ell(\hat{\beta}_R, \tilde{\eta}) - \ell(\hat{\beta}_R, \tilde{\eta})] \]

\[ \leq 2[\ell(\hat{\beta}_u, \hat{\eta}) - \ell(\hat{\beta}_R, \hat{\eta})] \]

\[ = LR \]

so the inequality \( LR \geq LM \) continues to hold for tests of nonlinear restrictions upon the coefficients of a linear regression.

For the W test, the unrestricted estimate will be \( \hat{\beta}_u \) as defined in §2.2, giving the statistic

\[ W = (\phi(\hat{\beta}_u))'[\tilde{F}'(X'\hat{\Omega}^{-1}X)^{-1}\tilde{F}]^{-1}[\phi(\hat{\beta}_u)] \]

where \( \tilde{F} = F(\hat{\beta}_u) \). With \( \phi(\hat{\beta}_u) \) nonlinear in \( \hat{\beta}_u \) it will not be possible to reduce the W statistic to the difference between quadratic forms in the residual vectors that enables it to be interpreted as coming from a conditional likelihood ratio procedure. In such a situation it is doubtful if a systematic relationship exists between W and the other statistics; it should therefore be possible to find cases in which \( W < LR \) or even \( W < LM \).

Similar complications arise when testing linear or nonlinear restrictions upon the parameters of nonlinear models, where no systematic
relationship can be found by the method used in §2.3. It would appear that the result is principally a feature of linearity of the model and the hypothesis.

2.7 Other Extensions

The conditions of independence between the coefficient parameters and those parameters upon which the covariance matrix depends are difficult to relax without losing the inequality result. With functional relationships between the two sets of parameters (e.g. in heteroscedastic models where the variance at each observation is related to the regression mean at that observation), the information matrix will be much more complicated than when the covariance matrix parameters are separate entities. In such a situation, the test statistics derived in §2.2 would not be correct and the interpretation of W and LM as conditional likelihood ratio procedures breaks down entirely.

It is also possible for the two sets of parameters \( \beta \) and \( \alpha \) to be statistically related even though the parameters of the covariance matrix, \( \alpha \), are not functions of the regression coefficients. When the information matrix has nonzero elements in the off-diagonal blocks connecting \( \beta \) with \( \alpha \), maximum likelihood estimators of \( \beta \) and \( \alpha \) will be asymptotically correlated, and the formulae obtained in §2.2 for the W and LM statistics will be inappropriate. The example given by Savin (1976) has one of the regressors being the first lag of the dependent variable when the disturbance is first-order autoregressive. Another common econometric specification which exhibits the same feature is the simultaneous equations system.

---

5 A problem with extensions to nonlinear regression models is ambiguity of the estimate of the information matrix for use in W and LM tests.
If the information matrix is not block diagonal between \( \beta \) and \( \alpha \), the formulae in §2.2 for estimation of \( \beta \) conditional upon an estimate of \( \Omega \) will still be correct, provided regressors are not contemporaneously correlated with the disturbances. (Simultaneous equation models are mentioned separately below.) However the formulae for the \( W \) and LM statistics will have to be modified; the correct forms will be

\[
W = (F' \hat{\beta}_u - \hat{\beta})'(F' I_{BB}^{-1} F)(F' \hat{\beta}_u - \hat{\beta})
\]

\[
LM = \lambda' F' I_{BB}^{-1} F \lambda
\]

where \( I_{BB} = (I_{BB}^{-1} - I_{BB})^{-1} \) from partitioned inversion of the information matrix. Since \( I_{BB}^{-1} \) is positive semidefinite when \( I \) is positive definite, and it was \( I_{BB}^{-1} = (X' \Omega^{-1} X)^{-1} \) that was used to obtain the conditional likelihood ratio interpretations in §2.3, for the correct statistics:

\[
W \leq (F' \hat{\beta}_u - \hat{\beta})'(F' I_{BB}^{-1} F)(F' \hat{\beta}_u - \hat{\beta})
\]

\[
= 2[\ell(\hat{\beta}_u, \hat{\Omega}) - \ell(\hat{\beta}_u, \hat{\Omega})]
\]

\[
> LR
\]

and

\[
LM \geq \lambda' F' I_{BB}^{-1} F \lambda
\]

\[
= 2[\ell(\hat{\beta}_R, \hat{\Omega}) - \ell(\hat{\beta}_R, \hat{\Omega})]
\]

\[
< LR
\]

Therefore no systematic relationship can be expected among the three testing criteria. Savin (1976, Section 8) gives empirical examples of both \( W > LR \) and \( W < LR \) when the proper \( W \) test is used in a case where coefficient and covariance matrix parameters are related via the
information matrix.

For the simultaneous equations model, suggested by Berndt and Savin (1977, p.1276) as a possible direction of extension of the inequality result, the situation is even more difficult. In this case, not only is the information matrix not block diagonal, but the simple linear representations of coefficient estimators conditional on covariance matrix estimates no longer hold. While it may be possible to define estimators which sufficiently fit the structure of §2.2 to be treated as iterated GLS estimation of regression models (e.g. 3SLS with iteration upon the covariance matrix but not the instruments and with the same reduced form predictions replacing endogenous regressors in restricted and unrestricted formulations of the model), it is doubtful that a search for systematic inequality relationships among the criteria would prove very fruitful.

2.8 Significance of the Inequality.

An implication of the inequality relationship, $W > LR > LM$ among the calculated values of the three statistics, is that typically there will exist certain significance levels at which the criteria will lead to conflicting inferences when they are referred to their common asymptotic $\chi^2$ distribution to set the critical value. However, a systematic inequality is not necessary for there to be conflict among the criteria; conflict is possible whenever alternative procedures that are only asymptotically equivalent are applied to finite samples.

Even if exact distributions were to be used, conflicting inferences can be given by alternative procedures in many situations. In the multivariate linear regression, for example, unless $M = 1$ the three statistics, $W$, $LR$ and $LM$ take different parts of the sample space as their critical region and so, even if the probability of Type I error is
the same with all three criteria, the decision from one given sample need not be the same in each case. When $M = 1$, giving the univariate linear model with scalar covariance matrix that was examined in §2.4, the three criteria are one-to-one transformations of each other. Hence in this case they would give exactly the same inference if exact distributions were to be used, although the numerical inequality relation holds among the statistics as they are defined. But this is a very special situation in which most of the asymptotic approximations, by which the three hypothesis testing criteria are related, hold exactly. In general, use of exact finite sample distributions to set significance levels would not resolve the conflict.

Therefore the systematic numerical inequality is neither necessary nor sufficient for there to be conflict among the criteria. The main significance of the relationship appears to be that a researcher, who is constrained (or elects) to use approximate significance points from the asymptotic $\chi^2$ distribution, can favour one outcome over another by a priori choice of the testing procedure to be employed. In those models for which the result holds, the $W$ statistic is more likely to lead to rejection of the null hypothesis and acceptance of the hypothesis is favoured by opting to use the LM criterion. Berndt and Savin (1977) consider this to be a "disturbing implication" of the inequality relationship in that "... if the researcher has subjective preferences regarding the truth of the null hypothesis, he can consult [the inequality] and judiciously choose the test procedure which is most likely to provide supporting evidence" [p.1275]. But the availability of alternative inference methods, which are equivalent only asymptotically and which can give quite different results in finite samples, allows such judicious choices to be made a posteriori anyway. When systematic relationships
(or similar general tendencies) can be established to hold between the outcomes of different methods that might be applicable in the same situation, then reported empirical results can be judged accordingly.

It has been suggested by Mizon (1977) that the inequality relationship may be a desirable one in situations where many hypotheses are to be tested. If the maintained model is very general and a number of sets of possible restrictions are to be tried, then the W procedure will be the most attractive computationally, as discussed in §1.6. When the inequality holds, the W test will also be the most stringent of the three testing procedures, and this reflects an appropriate treatment of the situation of trying many restrictions on a maintained general model. With tests of misspecification, where the maintained hypothesis is the most restrictive formulation and a number of possible generalizations of it are entertained, less stringent testing is argued to be desirable. The LM test which will be the easiest to apply in this situation will also be the least likely to reject the null hypothesis which represents the maintained specification.

It should be made clear that the inequality has no implications for the relative powers of the procedures in the classical sense. One statistic cannot be said to be more powerful than another simply because it is numerically larger and hence more likely to reject the hypothesis when a common critical value is employed. The inequality still holds when the hypothesis is correct (unless the hypothesis is exactly satisfied by the sample), indicating that the true sizes of the tests are different.
CHAPTER 3
THE LAGRANGE MULTIPLIER TEST AND ITS APPLICATION
TO MODEL SPECIFICATION*

3.1 Introduction

Many econometric procedures are such that their stochastic properties are readily susceptible to analysis only by asymptotic techniques. In the introductory remarks to Chapter 1, it was noted that there are three principles based on asymptotic theory for constructing tests of parametric hypotheses. While W and LR tests are frequently employed in econometric practice, the asymptotically equivalent LM principle has not been used, at least not explicitly, to the same degree.

It is true that a number of writers have drawn the econometrician's attention to the LM testing procedure: e.g., Byron (1970) tested individual restrictions within an imposed set by reference to their corresponding multipliers and Dhrymes et al. (1972) illustrated application of the procedure in the simple linear model. But it would appear that usage of the LM principle has been directed least to those situations in which it proves to be most useful.

In the process of validating an econometric model, various tests are usually performed to determine if a more general specification may be preferable. Such tests, called "diagnostic checks" by Box and Jenkins (1970) or "tests of misspecification" by Mizon (1977), would encompass all of the familiar procedures for detecting omitted explanatory variables, autocorrelation, heteroscedasticity, structural change etc. that are based upon the residuals from the fitted model. As was observed in §1.6 and at

* This chapter overlaps considerably with Breusch and Pagan (1979b).
the end of the previous chapter, it is in this situation of misspecification testing that the LM procedure will be most attractive for its computational simplicity. While W and LR tests could be used for diagnostic checking, they will be less attractive in practice because estimation of an alternative generalization typically involves complications such as nonlinearities or nonspherical disturbances. The LM procedure requires estimation of only the more restrictive specification and these estimates will already be available when diagnostic checking is contemplated. It will be seen that in many instances the LM statistic can be computed in a least squares regression using the residuals from the fitted model.

Analysis of residuals is an obvious and important approach to detecting misspecification of a statistical model. At one extreme would be the graphical methods with visual examination or other nonparametric techniques, e.g. as developed in a general context by Anscombe and Tukey (1963). For the linear regression model, many tests have been devised for specification errors with parametric representations. But sometimes the question remains whether, or in what sense, the criteria that are adopted are appropriate for the potential error. An alternative approach to developing diagnostic checks would be to use a formal hypothesis testing framework, and in econometric applications this would usually mean appealing to asymptotic theory for optimality properties. The LM principle is particularly useful as a guide to the development of new diagnostic tests and as a unifying principle for tests that have otherwise been considered in isolation.

Use of the LM procedure to construct misspecification tests requires that an alternative parametric model be specified in such a way that the current formulation can be obtained by imposing restrictions upon the parameters of the full model, typically by constraining certain parameters to
be zero. Thus the null hypothesis asserts the current specification to be the correct one. What are sometimes referred to as "nuisance parameters" because they are present but unknown under both null and alternative hypotheses will be the parameters of the maintained model. Application of the LM method need not, however, be confined to situations that are normally considered as being misspecification testing or diagnostic checking. Whenever it is easier to estimate the more restrictive model the LM procedure will be attractive for its computational economy relative to W or LR tests.

In the remainder of this chapter, further aspects of the LM test are discussed and econometric applications of it are developed. The framework of §3.2 and §3.3 is rather general, with alternative forms of the criterion being discussed in §3.2 and the connection with Neyman's C(α) tests being explored in §3.3. Applications relate to nonlinear regression models in §3.4, to the disturbance covariance matrix of linear models in §3.5 and §3.6, and in §3.7 consideration is given to a situation in which the information matrix is singular in the unrestricted model. Some other features of the LM test are noted in the concluding §3.8. Further applications which emphasize the role of the LM procedure as a framework for developing tests of misspecification are treated in more detail in Chapters 4 through 6.

3.2 Forms of the LM Statistic

The LM statistic for testing p hypothesized constraints on s parameters, formulated as \( \psi(\theta) = 0 \), was given in §1.3 both as a function of the Lagrange multipliers arising in constrained likelihood maximization and as the score test statistic,
Here \( \tilde{\theta} \) is the vector of estimates from maximizing the log-likelihood \( \ell(\theta) \) subject to \( \phi(\theta) = 0 \), and the tilde indicates that the score vector \( d(\theta) = \partial\ell/\partial\theta(\theta) \) and the information matrix \( I(\theta) = -E[\partial^2\ell/\partial\theta\partial\theta'(\theta)] \) are evaluated at \( \theta = \tilde{\theta} \). It was noted that variations on (1) with the same asymptotic properties are made possible by using matrices other than \( \tilde{I} \) as the matrix in the quadratic form. One alternative would be to use the negative of the Hessian matrix of second derivatives or, when the log-likelihood is essentially a sum of squares function, the weighting matrix of the Gauss-Newton algorithm could be used instead. This latter choice will be seen in §3.4 to give a regression solution for computing the LM statistic in many problems, particularly when testing for misspecification errors in regression models.

There is an interesting interpretation of the LM test as a "pseudo-Wald test", which may give a useful indirect method for computing the criterion in some situations. Observe that the method of scoring, applied to the unrestricted generalization but using the restricted estimates \( \tilde{\theta} \) as initial parameter values, would yield at the first round:

\[
\hat{\theta}^* = \tilde{\theta} + \tilde{I}^{-1}d(\tilde{\theta}) .
\]  

(2)

Rearranging and substituting in (1),

\[
d(\tilde{\theta}) = \tilde{I}(\hat{\theta}^* - \tilde{\theta})
\]

\[
LM = (\hat{\theta}^* - \tilde{\theta})'\tilde{I}(\hat{\theta}^* - \tilde{\theta}) .
\]  

(3)

This is the quantity that would be used in a Wald-type test for the "hypothesis" \( \theta = \tilde{\theta} \), using the results from the first round of scoring as estimates of the parameters of the full model. Note, however, that it would not be a proper \( W \) test of \( \phi(\theta) = 0 \), not even the \( W \) test using
efficient two-step estimators that was discussed in §1.9. The $W$ criterion for testing $\phi(\delta) = 0$ would be a quadratic form in $\phi(\hat{\delta})$ with $\hat{\delta}$ being the unrestricted maximum likelihood estimates or, in the case of two-step estimation, with estimates having the same asymptotic distribution when the alternative hypothesis is correct. But the "estimates" $\hat{\delta}^*$ given by (2) will in general have none of the usual desirable properties: if the null hypothesis is true they will be consistent but inefficient relative to the restricted estimates $\hat{\delta}$, otherwise they will be inconsistent because the initial values $\hat{\delta}$ are constrained by false restrictions. While the construction in (3) is rather curious when viewed as a test on $(\hat{\delta}^* - \hat{\delta})$, its derivation from the LM statistic shows that it would be valid as a criterion for testing $\phi(\delta) = 0$.

This interpretation of the LM criterion could be useful in some situations. Suppose that a numerical algorithm has been set up to obtain maximum likelihood estimates for the alternative model and, as would typically be done, the available estimates for the more restrictive specification are to be used as initial values for the parameters. If the algorithm is based upon scoring, Newton-Raphson, Gauss-Newton or any adaptation of these methods that updates the $m$'th approximate solution by

$$\delta^{m+1} = \delta^m + H^{-1}d(\delta^m)$$

where $[I^{-1}H] \overset{P}{\rightarrow} I_s$, then the LM test can be performed after the first round to see whether estimation of the alternative model (i.e. continued iteration) is likely to be worthwhile. In cases where it is difficult to obtain derivatives analytically, such as the Goldfeld-Quandt (1973) switching regressions model or the Rosenberg (1973) varying coefficient specification, the indirect approach could be advantageous. Hypotheses could be tested by the indirect LM method using an estimation algorithm.
which employs numerical derivatives, but without continuing iterations beyond the first.\footnote{When numerical derivatives are to be used, some restrictions are placed (conceptually) on the method for determining step length if asymptotic equivalence is to be preserved. [See Sargan (1975, Sect. 3).]}

A very similar approach, of course, would be to use one round of the same numerical algorithm to obtain consistent and efficient two-step estimates and to use these in a \( W \) test. But this would require that the initial estimates in the two-step procedure be consistent when the alternative hypothesis is correct, and the additional computations for this first step may indicate that the indirect LM test is preferable for its relative simplicity.

As a test for misspecification, the LM procedure will generally be formulated for the situation where the null hypothesis asserts some of the parameters to have specific values (typically zeros). Partitioning \( \theta' = (\theta_1', \theta_2') \), the null hypothesis is \( \theta_1 = \theta_{10} \) and (1) becomes

\[
\text{LM} = (d_{1}(\theta_{10}', \hat{\theta}_2'))'I_{11}^{-1}[d_{1}(\theta_{10}', \hat{\theta}_2')]
\]

where \( d_{1}(\theta_{10}', \hat{\theta}_2') \) is the subvector of \( d(\theta) \) and \( I_{11}^{-1} \) is the submatrix of \( I^{-1} \) corresponding to \( \theta_1 \) parameters. The other component of \( d(\theta) \) is \( d_{2}(\theta_{10}', \hat{\theta}_2') \) which is zero from the first-order conditions on \( \hat{\theta}_2 \) as the maximum likelihood estimates under the null hypothesis.

One round of scoring starting from initial estimates \( \hat{\theta}' = (\hat{\theta}_{10}', \hat{\theta}_2') \) gives in this case:

\[
\begin{bmatrix}
\hat{\theta}_1^* \\
\hat{\theta}_2^*
\end{bmatrix} =
\begin{bmatrix}
\theta_{10} \\
\hat{\theta}_2
\end{bmatrix} +
\begin{bmatrix}
\hat{I}_{11} & \hat{I}_{12} \\
\hat{I}_{21} & \hat{I}_{22}
\end{bmatrix}
\begin{bmatrix}
d_{1}(\theta_{10}', \hat{\theta}_2') \\
0
\end{bmatrix}
\]

so that
and, substituting in (4),

$$\text{LM} = (\hat{\theta}_1^* - \theta_{10})' [\hat{I}^{11}]^{-1} (\hat{\theta}_1^* - \theta_{10}).$$

(5)

This appears even more like a $W$ test criterion than expression (3) above for the case of general restrictions, but it should be distinguished from a proper $W$ statistic for the same reasons. Again, the LM statistic can be computed indirectly when analytical derivation is difficult.

### 3.3 The $C(\alpha)$ Statistic

When the null hypothesis specifies particular values for a subset of the parameters there is another procedure which has been proposed, using a criterion which is closely related to the LM statistic. This is the so-called $C(\alpha)$ test of Neyman (1959) which was extended to the case of multiple constraints by Buhler and Puri (1966). Whereas the LM test is based upon constrained maximum likelihood estimation, the $C(\alpha)$ procedure requires estimates of the nuisance parameters which are only root-$n$ consistent and not necessarily efficient (assuming root-$n$ norming to be appropriate). The $C(\alpha)$ criterion is

$$C(\alpha) = [d_1(\theta_{10}, \bar{\theta}_2) - \bar{T}_{12}^{-1} d_2(\theta_{10}, \bar{\theta}_2)]' \bar{I}^{11} [d_1(\theta_{10}, \bar{\theta}_2) - \bar{T}_{12}^{-1} d_2(\theta_{10}, \bar{\theta}_2)]$$

(6)

where $\bar{\theta}_2$ need be only "locally root-$n$ consistent", i.e.

$$\sqrt{n}(\bar{\theta}_2 - \theta_2) \sim N_p(0, 1)$$

under the null hypothesis and for local alternatives, and where $\bar{T} = I(\bar{\theta})$. Although asymptotic equivalence with the $W$ and LR statistics was implicit earlier, a clear demonstration of this may be found in Moran (1970).

---

2 A third method also closely related to the LM test was used by Durbin (1970) and is considered in Chapter 4 within the context for which it was developed.
Rearranging (6) using the partitioned inverse of $\bar{I}$,

$$C(\alpha) = [d(\theta_{10}, \bar{\theta}_2)]' \begin{bmatrix} \bar{I}^{11} & \bar{I}^{12} \\ \bar{I}^{21} & \bar{I}^{22} - \bar{I}^{-1} \\ \end{bmatrix} [d(\theta_{10}, \bar{\theta}_2)]$$

$$= [d(\theta_{10}, \bar{\theta}_2)]' \bar{I}^{-1} [d(\theta_{10}, \bar{\theta}_2)] - [d_2(\theta_{10}, \bar{\theta}_2)]' \bar{I}^{-1} [d_2(\theta_{10}, \bar{\theta}_2)]. \quad (7)$$

Compared with the LM formula (1), the $C(\alpha)$ criterion includes an additional term which reduces the statistic to account for $d_2(\theta_{10}, \bar{\theta}_2) \neq 0$ when the estimates $\bar{\theta}_2$ are merely consistent and not constrained maximum likelihood. Alternatively from (6), the correction can be thought of as focusing on $d_1(\theta_{10}, \bar{\theta}_2)$ conditional upon the value of $d_2(\theta_{10}, \bar{\theta}_2)$.

Indirect computation of the $C(\alpha)$ statistic is also possible using one round of a numerical estimation algorithm, this time with $(\theta_{10}, \bar{\theta}_2)$ as the initial estimates, extending the interpretation of the LM test as a pseudo-Wald test to the situation where initial estimates of the restricted model are merely consistent under the null hypothesis.

Although not directly related to the Lagrange multipliers arising in constrained likelihood maximization, the $C(\alpha)$ statistic has a very similar construction. If constrained maximum likelihood estimates were to be used for the nuisance parameters $\theta_2$ or if the information matrix were to be block diagonal such that $I_{12} = I_{21} = 0$, then $C(\alpha)$ and LM methods would coincide. The $C(\alpha)$ procedure is more general than the LM test in that it makes weaker demands on the properties of estimates of the restricted model, but reparameterization would be required for $C(\alpha)$ to be applicable to tests of general constraints. One advantage

---

3 It was noted in §1.9 that the LM test could be used unmodified when $\sqrt{n}(\bar{\theta}_2 - \bar{\theta}_2) = o_p(1)$; in this case it can be shown that the correction term in (7) is asymptotically negligible.
of recognizing the connection between C(α) and LM tests is that results in the literature relating to the former, such as properties of the tests when parameters lie on boundaries, can be seen to apply equally to the latter.

3.4 Nonlinear Regression Models

There is a class of problems in which the model may be written as the nonlinear regression,

\[ y_t = g(z_t; \theta) + \epsilon_t \]  \hspace{1cm} (8)

where \( \epsilon_t \sim \text{NID}(0, \sigma^2) \) independently of \( g_t \equiv g(z_t; \theta) \) for \( t = 1, \ldots, n \). Partitioning the \( s \) parameters in \( \theta \) as \( \theta' = (\theta_1', \theta_2') \), an hypothesis involving \( p \) constraints as \( \theta_1 = \theta_{10} \) is to be tested. One common way in which this situation may arise is where a more restrictive model, represented by (8) with \( \theta_1 = \theta_{10} \), has been estimated and it is then suspected that the more general formulation may be appropriate, i.e. as a test of misspecification.

Defining the vector of disturbances as \( \epsilon' = (\epsilon_1, \ldots, \epsilon_n)' \), the log-likelihood for the full model is

\[ \ell(\theta, \sigma^2) = -\frac{1}{2} n \log (2\pi) - \frac{1}{2} n \log \sigma^2 - \frac{1}{2} \sigma^{-2} \epsilon' \epsilon \]  \hspace{1cm} (9)

where \( \epsilon \) is the function of the data and the unknown parameters given by rearranging (8). In some cases, (9) may be only an asymptotic approximation to the log-likelihood; e.g., if \( z_t \) contains lagged values of the dependent variable (or of the disturbance as in example (ii) below) and initial conditions are ignored on the assumption that asymptotic properties of estimators and tests will be unaffected by such approximations. Using the notation,
\[ G_t(0) = \frac{\partial \varepsilon_t}{\partial \theta(0)} = - \frac{\partial \varepsilon_t}{\partial \theta(0)} \]

and

\[ G = (G_1, \ldots, G_n)'
\]

the score vector component for \( \theta \) is

\[ d_\theta(\theta, \sigma^2) = \sigma^{-2} G' \varepsilon \]

and the required component of the information matrix is

\[ I_{\theta \theta} = \sigma^{-2} E(G'G) . \]

The latter follows from

\[ \frac{\partial^2 \ell}{\partial \theta \partial \theta'} = -\sigma^{-2} [G'G - \sum_t (\sigma^{-2} \frac{\partial \varepsilon_t}{\partial \theta} \varepsilon_t)] \]

with the second term having zero expectation because of the assumption of independence between \( \varepsilon_t \) and \( \varepsilon_t' \). Also, the information matrix will be block diagonal between \( \theta \) and \( \sigma^2 \) components because

\[ \frac{\partial^2 \ell}{\partial \theta \partial \sigma^2} = -\sigma^{-4} G' \varepsilon = -\sigma^{-4} \sum_t G_t \varepsilon_t \]

has zero expectation. Maximum likelihood estimates of the restricted model, which are the estimates of the full model under the constraint of the null hypothesis, are denoted by \( \tilde{\theta}' = (\tilde{\theta}_1', \tilde{\theta}_2') \) and \( \tilde{\sigma}^2 = n^{-1} \tilde{\varepsilon}' \tilde{\varepsilon} \), where

\[ \tilde{\varepsilon}_t = y_t - g(z_t; \tilde{\theta}) . \]

With the information matrix being block diagonal between \( \theta \) and \( \sigma^2 \) parameters, the LM statistic for testing \( H_0: \theta_1 = \theta_10 \) is

\[ LM = [d_\theta(\tilde{\theta}, \tilde{\sigma}^2)]' I_{\theta \theta}^{-1} [d_\theta(\tilde{\theta}, \tilde{\sigma}^2)] \]
where an estimate is required for \( I_{\theta\theta} = \sigma^{-2} E(G'G) \). Two sensible procedures (which would coincide if the \( z_t \) explanatory variables were non-stochastic) are suggested for this: one is to take the expectation and then replace any unknown parameters by their estimates in the constrained model; the other is to take

\[
\tilde{I}_{\theta\theta} = \tilde{\sigma}^{-2} \tilde{G}'\tilde{G}
\]

where \( \tilde{G} = G(\tilde{\theta}) \) as in the weighting matrix for the Gauss-Newton algorithm.

Using the second suggested choice for \( \tilde{I}_{\theta\theta} \), the criterion becomes

\[
\text{LM} = \tilde{\sigma}^{-2} \tilde{\epsilon}' \tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{\epsilon}
\]

\[
= n \tilde{\epsilon}' \tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{\epsilon}/\tilde{\epsilon}'\tilde{\epsilon}
\]

(10)

and the LM test using the asymptotic distribution would reject the current restricted model in favour of the generalization when the criterion exceeds the critical value taken from the \( \chi^2(p) \) distribution. Observe that the LM statistic for testing whether an explanatory variable (or combination of variables) has been omitted from the current specification of a nonlinear regression model can be computed in a simple linear regression. After fitting the model in its more restrictive mode, construct the quantities \( \tilde{G}_t \) using the form of the alternative hypothesis generalization and the restricted parameter estimates. Then the statistic can be obtained as \( n R^2 \), where \( R^2 \) is the usual coefficient of determination in the linear regression of \( \tilde{\epsilon}_t \) upon \( \tilde{G}_t \).

This regression strategy is, of course, an application of the indirect principle for computing the LM statistic that was given in §3.2 above. One round of Gauss-Newton, starting from initial estimates \( \tilde{\theta} \), would give

\[
\tilde{\theta}^* = \tilde{\theta} + (\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{\epsilon}.
\]
as a regression solution to updating the estimates. Then the indirect LM principle using (3) with \( \hat{I}_{\theta \hat{\theta}} = \hat{\sigma}^{-2} \hat{G}'\hat{G} \) would yield the criterion as in (10). Alternatively, partitioning \( \hat{G} = (\hat{G}_1 : \hat{G}_2) \) conformably with \( \theta' = (\theta'_1, \theta'_2) \) and noting that \( \hat{G}_2^T \hat{\varepsilon} = 0 \) from the first-order conditions for constrained likelihood maximization, the criterion from either (5) or (10) becomes

\[
LM = n \hat{\varepsilon}' \hat{G}_1 [\hat{G}_1^T \hat{G}_1 - \hat{G}_1^T \hat{G}_2 (\hat{G}_2^T \hat{G}_2)^{-1} \hat{G}_2^T \hat{G}_1]^{-1} \hat{G}_1^T \hat{\varepsilon} / \hat{\varepsilon}' \hat{\varepsilon}.
\]

These and other regression forms of the LM statistic are considered in more detail for a particular application in §4.6.

If the alternative hypothesis is to add a nonlinear complication to a linear model, the LM test can be performed by estimating two linear regressions only, without estimating the nonlinear specification. Suppose that the alternative generalization is

\[
y_t = x_t' \beta + g_2(z_t; \theta) + \varepsilon_t
\]

where the hypothesis is \( \theta = \theta_o \) such that \( g_2(z_t; \theta_o) = 0 \) for all \( t = 1, \ldots, n \). The LM test for this hypothesis would be performed by estimating the linear regression of \( y_t \) upon \( x_t \) by OLS to obtain the residuals \( \hat{\varepsilon}_t \) and then taking the \( \chi^2 \) criterion as \( n \hat{\varepsilon}' \hat{\varepsilon} \) in a second regression of \( \hat{\varepsilon}_t \) upon \( x_t \) and \( \hat{G}_2 = \partial g_2 / \partial \theta(\theta_o) \). In this situation, if \( x_t \) and \( z_t \) in (11) are nonstochastic, exact significance points may be taken from the F distribution, because the LM criterion is a one-to-one transformation of the usual F statistic for testing the significance of the \( \gamma \) coefficients in the linear model

\[
y_t = x_t' \beta + \hat{G}_2^T \gamma + \varepsilon_t
\]

which has nonstochastic regressors. For the corresponding W and LR criteria, on the other hand, iterative calculations would be required to
solve nonlinear normal equations for full estimation of the unrestricted model, so these statistics would generally have intractible finite sample distributions.\footnote{Gallant (1975) gives some evidence from simulation experiments that using the transformation including "degrees of freedom" adjustment that gives the \( F \) statistic in the linear case may provide a better guide to the distribution of the \( LR \) statistic in nonlinear regression models than the asymptotic \( \chi^2 \) approximation.}

A more general regression situation would have nonspherical disturbances with \( \varepsilon \sim N(0, \Omega) \), where \( \Omega \) is either known or it is parameterized so that the unknown parameters upon which it depends can be estimated. Denoting by \( \tilde{\Omega} \) the estimate of \( \Omega \) from the more restrictive model under the null hypothesis, the \( LM \) statistic would be

\[
LM = \varepsilon' \tilde{\Omega}^{-1} G (G' \tilde{\Omega}^{-1} G)^{-1} G' \tilde{\Omega}^{-1} \varepsilon
\]

provided the \( \theta \) parameters are not related to those upon which \( \Omega \) depends, either functionally or through the information matrix.

Some examples will now be considered to illustrate application of the \( LM \) test in nonlinear regression models.

(i) Testing for a Liquidity Trap

Konstas and Khouja (1969) tested for a liquidity trap in a demand for money function but, as Spitzer (1976) observed, their algorithm failed to maximize the likelihood so that their \( LR \) test was misleading. The equation estimated was

\[
M_t = \gamma Y_t + \beta (r_t - \alpha)^{-1} + \varepsilon_t
\]

with \( M_t, Y_t \) and \( r_t \) being money demand, income and the rate of interest respectively, and attention centred upon whether \( \alpha = 0 \), i.e. whether there was a liquidity trap. With \( \theta' = (\alpha, \beta, \gamma) \), the \( LM \) statistic can be
computed in two steps:

(a) regress \( M_t \) upon \( Y_t \) and \( r_t^{-1} \) to obtain estimates \( \hat{Y}, \hat{\beta} \) and residuals \( \hat{\varepsilon}_t \);

(b) form \( G'_t = [\beta (r_t - \alpha)^{-2}, (r_t - \alpha)^{-1}, Y_t] \) giving \( \hat{G}'_t = [\hat{\beta} r_t^{-2}, r_t^{-1}, Y_t] \) and regress \( \hat{\varepsilon}_t \) upon \( r_t^{-2}, r_t^{-1} \) and \( Y_t \), computing the coefficient of determination in this second regression. The value of \( LM(nR^2) = 11.47 \) would lead to rejection of the null hypothesis at all of the usual significance levels when taken as a \( \chi^2(1) \), thus agreeing with Spitzer's LR test.

(ii) **Testing for Autocorrelation**

As an example of applying the LM principle using the first suggested choice for estimating the information matrix instead of the regression solution, testing for autocorrelation is considered. Suppose the full model is

\[
\begin{align*}
y_t &= x'_t \beta + u_t \\
u_t &= \rho u_{t-1} + \varepsilon_t
\end{align*}
\]

where \( x_t \) is exogenous and the null hypothesis is \( \rho = 0 \). The unrestricted model can be written as the nonlinear regression

\[
y_t = \rho y_{t-1} + x'_t \beta - \rho x'_{t-1} \beta + \varepsilon_t
\]

giving \( G'_t = [(y_{t-1} - x'_t \beta), (x_t - \rho x_{t-1})'] \) for \( \theta' = (\rho, \beta') \).

In this case it is not difficult to show that \( I_{\rho \beta} = I'_{\beta \rho} = 0 \), so that

---

5 Recently, Engle (1978) has observed that the Konstas-Khouja specification of the money demand function is subject to considerable autocorrelation, so that a formulation of the LM criterion as in (13) would be more appropriate.
only the components of the score vector and information matrix corresponding to $\rho$ are required. They are

\[
d_{\rho}(\theta, \sigma^2) = \sigma^{-2} \sum_{t} u_{t-1} \varepsilon_{t}
\]

\[
I_{\rho \rho} = n/(1-\rho^2).
\]

Under the null hypothesis $\rho = 0$, maximum likelihood estimates are given by applying OLS to (14) to get coefficient estimates $\hat{\beta}$, residuals $\tilde{\varepsilon}_t = \tilde{u}_t = y_t - x_t'\hat{\beta}$ and variance estimate $\hat{\sigma}^2 = n^{-1} \sum_{t} \tilde{u}_t^2$. Then

\[
LM = [\hat{\sigma}^{-2} \sum_{t} \tilde{u}_{t-1} \tilde{u}_t]'[n]^{-1}[\hat{\sigma}^{-2} \sum_{t} \tilde{u}_{t-1} \tilde{u}_t]
\]

\[
= n \left( \sum_{t} \tilde{u}_{t-1} \tilde{u}_t \right)^2 / \left( \sum_{t} \tilde{u}_t^2 \right)^2
\]

\[
= n[2-DW]^2
\]

where DW is the usual Durbin-Watson statistic for diagnosing first-order autocorrelation using OLS residuals. Thus the DW statistic can be obtained from the LM criterion for this problem; the transformation from LM to DW allows a sign to be attached to the statistic to permit testing of a one-sided alternative hypothesis.

(iii) Testing Functional Form

Andrews (1971) developed a simple and exact test for the hypothesis $H_0: \lambda = \lambda_0$, where $\lambda$ is the parameter of the transformation $y_t \rightarrow y_t^{(\lambda)}$ which gives the dependent variable for the regression

\[
y_t^{(\lambda)} = x_t'\beta + \varepsilon_t
\]

with $\varepsilon_t \sim NID(0, \sigma^2)$ for $t = 1, \ldots, n$. The test is motivated by a local linearization of the dependent variable,
\[ y_t^{(\lambda)} = y_t^{(\lambda_0)} + v_t^{(\lambda - \lambda_0)} \]

where \( v_t = \partial y_t^{(\lambda)}/\partial \lambda \), evaluated at \( \lambda = \lambda_0 \). Writing all \( n \) observations in obvious vector notation, the linearized model is

\[ y^{(\lambda_0)} = X\beta + V(\lambda_0 - \lambda) + \varepsilon. \]

Andrews' test is performed in two steps:

(a) regress \( y^{(\lambda_0)} \) upon \( X \) to obtain residuals \( \tilde{\varepsilon} \) and predictions \( \tilde{y}^{(\lambda_0)} \);

(b) form \( \tilde{U} = Q\tilde{V} \) where \( Q = I - X(X'X)^{-1}X' \) and \( \tilde{V} \) is \( V \) evaluated at \( \tilde{y}^{(\lambda_0)} \), and then test the overall significance of the regression of \( \tilde{\varepsilon} \) upon \( \tilde{U} \) using the usual \( F \) statistic.

Now, the \( F \) criterion for testing the overall significance of the regression of \( \tilde{\varepsilon} \) upon \( \tilde{U} \) will be exactly the same as that for testing the significance of \( \gamma \) coefficients in

\[ y^{(\lambda_0)} = X\beta + V\gamma + \varepsilon. \]

It would appear that Andrews' test can be viewed as a regression solution to computing the LM statistic where exact significance points can be applied because the model is linear under the null hypothesis and the regressors of the second regression are nonstochastic as in (12) above. However there is a complication. The log-likelihood is given from Box and Cox (1964, p.215) as

\[ \ell(\lambda, \beta, \sigma^2) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} n \log \sigma^2 - \frac{1}{2} \sigma^{-2} e'\varepsilon + \sum \log|\partial y_t^{(\lambda)}/\partial y_t| \]

where \( \varepsilon = y^{(\lambda)} - X\beta \), so that there is an additional Jacobian term.

In the particular case of the principal member of the family of
transformations proposed by Box and Cox, viz.

\[
y_{t}^{(\lambda)} = \begin{cases} 
(y_{t}^{\lambda} - 1)/\lambda & \text{for } \lambda \neq 0 \\
\log y_{t} & \text{for } \lambda = 0
\end{cases}
\]

the extra term is

\[
\sum \log |\frac{\partial y_{t}^{(\lambda)}}{\partial y_{t}}| = (\lambda-1) \sum \log y_{t}
\]

giving the component of the score vector corresponding to \( \lambda \) as

\[
d_{\lambda}(\lambda, \beta, \sigma^2) = -\sigma^{-2} (\partial \epsilon/\partial \lambda) \epsilon + \sum \log y_{t}
\]

\[
= -\sigma^{-2} \psi' \epsilon + \sum \log y_{t}.
\]

Thus the LM test, which would take as its critical region large (absolute) values of \( d_{\lambda}(\lambda, \beta, \sigma^2) \), differs from Andrews' test which rejects for large values of \( \psi' \epsilon \) and ignores the contribution from the Jacobian. However, as observed by Box and Cox (1964, p.216), normalizing the transformation by the geometric mean of the \( y_{t} \) observations removes the separate Jacobian term. Unless the dependent variable data is geometrically mean corrected (so that \( \sum \log y_{t} = 0 \)), the test proposed by Andrews will differ from the LM test although it bears a close resemblance to it.

### 3.5 Testing for Nonspherical Disturbances

Many aspects of potential misspecification in regression models focus on the covariance matrix of the disturbances. Suppose that the linear model with \( n \) observations,

\[
y = X\beta + \epsilon
\]
where $\varepsilon \sim N(0, \Omega)$, has been estimated by OLS on the assumption that $\Omega = \sigma^2 I_n$ but then it is suspected that some other specification, say $\Omega = \Omega(\theta)$, may be more appropriate. If $\theta$ is a $(p+1)$-vector of unknown parameters which are in principle estimable, the null hypothesis specification $\Omega = \sigma^2 I_n$ will usually correspond to imposing $p$ constraints on the general model as $\theta_1 = 0$ in the partitioning $\theta' = (\theta_1', \theta_2')$ with $\theta_2 = \sigma^2$.

A number of situations which permit differing treatments can be distinguished according to the relationships between the sets of parameters $\theta_1$, $\theta_2(= \sigma^2)$ and $\beta$ via the information matrix. In some models there will be no information links between $\theta_1$ on the one hand and $(\beta, \sigma^2)$ on the other; this would be the case if the previous example of testing for autocorrelation in regression models with purely exogenous regressors was cast explicitly in the framework of testing for nonspherical disturbances. In other situations, $\theta_1$ might be related to some elements in $\beta$ but be unconnected with $\sigma^2$; testing for autocorrelation in models with lagged dependent variables as regressors would fit this category. Another common situation is where $\theta_1$ and $\sigma^2$ are both unrelated to $\beta$, but $\theta_1$ may be related to $\sigma^2$ in the sense that the corresponding off-diagonal elements of the information matrix are nonzero; this is the class of specifications to be considered here.

For the full unrestricted model (15), the log-likelihood is

$$\ell(\theta, \beta) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \log|\Omega| - \frac{1}{2} \varepsilon' \Omega^{-1} \varepsilon$$

where $\varepsilon = y - X\beta$. To test the hypothesis $\theta_1 = 0$, the following quantities are required to form the LM statistic:

$$d_0(\theta, \beta) = \frac{1}{2} A'(\Omega^{-1} \theta \Omega^{-1}) \text{vec}(\varepsilon \varepsilon' - \Omega)$$

$$I_{\theta\theta} = \frac{1}{2} A'(\Omega^{-1} \theta \Omega^{-1})A$$
where \( \tilde{A} \) is the \( n^2 \times (p+1) \) matrix,

\[
\tilde{A} = A(\theta) = (3 \text{ vec } \Omega / \partial \theta)' .
\]

Under the constraint of the null hypothesis, maximum likelihood estimates will be given by applying OLS to (15) to give \( \tilde{\beta} \) and \( \tilde{\sigma}^2 = n^{-1} \tilde{e}' \tilde{e} . \) Evaluating the relevant parts of the score vector and information matrix at the restricted estimates gives

\[
d_{\theta} (\tilde{\theta}, \tilde{\theta}) = \frac{1}{2} \tilde{\sigma}^{-2} \tilde{A}' \text{ vec} (\tilde{\sigma}^{-2} \tilde{e}' - I_n) \\
I_{\theta \theta} = \frac{1}{2} \tilde{\sigma}^{-4} \tilde{A}' \tilde{A}
\]

where \( \tilde{\theta}' = (0', \tilde{\sigma}^2) \) so that \( \Omega(\tilde{\theta}) = \tilde{\sigma}^2 I_n . \) Putting together the quadratic form (4) that constitutes the LM statistic then gives

\[
LM = \frac{1}{2} [\text{ vec}(\tilde{\sigma}^{-2} \tilde{e} \tilde{e}' - I_n)]' \tilde{A}(\tilde{A}' \tilde{A})^{-1} \tilde{A}' [\text{ vec}(\tilde{\sigma}^{-2} \tilde{e} \tilde{e}' - I_n)]
\]

and this quantity could be computed as one half of the explained sum of squares in a regression with \( \text{ vec}(\tilde{\sigma}^{-2} \tilde{e} \tilde{e}' - I_n) \) as the dependent variable and \( \tilde{A} \) as the regressor set.

While all of the familiar parametric models of nonspherical disturbances could be handled in this way, it will frequently be the case that using (18) as a regression quantity would be unnecessarily cumbersome. If the matrix \( A \) is sparse, it may be more straightforward to focus directly on the parameters under test rather than computing a regression in \( n^2 \) "observations". For example, testing for autocorrelation in dynamic models is handled in Chapter 4 as a nonlinear regression problem, while the heteroscedasticity test of Chapter 6 is developed directly from simpler expressions for the required components of the score vector and information matrix. In both cases a regression solution in \( n \) observations is given for computation of the LM statistic.

As a more useful application of the general method, consider the error
components model of Balestra and Nerlove (1966) for N individuals observed over T time periods:

\[ y_{it} = X_{it} \beta + u_{it} \]  
\[ u_{it} = \nu_i + \lambda_t + \nu_{it} \]

where \( \nu_i, \lambda_t \) and \( \nu_{it} \) are mutually independent, distributed as

\[ \nu_i \sim \text{NID}(0, \sigma^2_\nu), \quad \lambda_t \sim \text{NID}(0, \sigma^2_\lambda), \quad \nu_{it} \sim \text{NID}(0, \sigma^2_v) \]

for \( i = 1, \ldots, N \) and \( t = 1, \ldots, T \). Following Nerlove (1971), equations (19) and (20) may be written more compactly in matrix form using

\[ i = (1, \ldots, 1)' \]  
\[ y = XB + u \]  

where

\[ u \sim \text{N}(0, \Omega) \]

\[ \Omega = \sigma^2_v I_{NT} + \sigma^2_\nu (I_N \otimes I_T) + \sigma^2_\lambda (I_N \otimes \theta) \]

with the subscripts on \( i_N \) and \( i_T \) denoting the dimension of the vector.

The null hypothesis that \( \Omega = \sigma^2_v I_{NT} \) permits efficient estimation of (21) by OLS with simple aggregation over time and individuals to give \( n = NT \) observations. This would be given by the restrictions

\( \sigma^2_\nu = \sigma^2_\lambda = 0 \) in the general specification, with the alternative hypothesis being \( \sigma^2_\nu > 0 \) and \( \sigma^2_\lambda > 0 \) with at least one strict inequality. Maddala (1971) discusses some of the problems inherent in maximum likelihood estimation of the full model, including multiple maxima of the likelihood and interpretation of boundary solutions. Furthermore, estimates obtained by maximizing the likelihood over the unrestricted parameter space, including \( \sigma^2_\nu > 0 \) and \( \sigma^2_\lambda > 0 \), cannot be asymptotically normal under the
null hypothesis which constrains $\sigma^2_\mu$ and $\sigma^2_\lambda$ to be on a boundary. The results of Chant (1974) suggest that, while the LM test remains unaffected by the boundary value hypothesis, corresponding W and LR tests will require special treatment because more than one parameter is specified by the hypothesis.

With $\theta' = (\sigma^2_\mu, \sigma^2_\lambda, \sigma^2_\nu)$, the information matrix will be block diagonal between $\theta$ and $\beta$ but, as will be seen, all $\theta$ derivatives will be required even though only a subset of $\theta$ is specified by the hypothesis. The matrix $A$ defined previously is in this case,

$$ A = (\partial \text{vec } \Omega/\partial \theta)' $$

$$ = \begin{bmatrix} \left( \frac{\partial \text{vec } \Omega}{\partial \sigma^2_\mu} \right)' : \left( \frac{\partial \text{vec } \Omega}{\partial \sigma^2_\lambda} \right)' : \left( \frac{\partial \text{vec } \Omega}{\partial \sigma^2_\nu} \right)' \end{bmatrix} $$

$$ = [\text{vec}(I_N \otimes \hat{\beta} \hat{\beta}') : \text{vec}(\hat{\beta} \hat{\beta}' \otimes I_T) : \text{vec } I_{NT}] .$$

Denoting OLS residuals from (21) as $\tilde{u}$ and the corresponding residual variance estimate (i.e. the estimate of $\sigma^2_\nu$ in the model constrained by the null hypothesis) as $\tilde{\sigma}^2 = (NT)^{-1}\tilde{u}'\tilde{u}$, the LM statistic would be given from (18) as one half of the explained sum of squares in the regression of $\text{vec}(\tilde{\sigma}^{-2} \tilde{u}\tilde{u}' - I_{NT})$ upon $\text{vec}(I_N \otimes \hat{\beta} \hat{\beta}')$, $\text{vec}(\hat{\beta} \hat{\beta}' \otimes I_T)$ and $\text{vec } I_{NT}$.\(^6\)

As an alternative to this regression with $(NT)^2$ observations, observe that

$$ A'A = NT \begin{bmatrix} T & 1 & 1 \\ 1 & N & 1 \\ 1 & 1 & 1 \end{bmatrix} $$

\(^6\) Alternatively $\text{vec}(\tilde{\sigma}^{-2} \tilde{u}\tilde{u}')$ could be used as the dependent variable and $NT$ subtracted from the resulting explained sum of squares.
\[ A' \text{vec}(\sigma^{-2} \tilde{u} \tilde{u}' - I_{NT}) = \begin{bmatrix}
\sigma^{-2} \tilde{u}'(I_N \otimes \hat{\lambda}_T') \tilde{u} - \text{tr}(I_N \otimes \hat{\lambda}_T') \\
\sigma^{-2} \tilde{u}'(\hat{\lambda}_N' \otimes I_T) \tilde{u} - \text{tr}(\hat{\lambda}_N' \otimes I_T) \\
\sigma^{-2} \tilde{u}' \tilde{u} - \text{tr}(I_{NT})
\end{bmatrix} \]

\[ = \sigma^{-2} \begin{bmatrix}
\tilde{u}'(I_N \otimes \hat{\lambda}_T' - I_{NT}) \tilde{u} \\
\tilde{u}'(\hat{\lambda}_N' \otimes I_T - I_{NT}) \tilde{u} \\
0
\end{bmatrix} \]

giving the LM statistic from (18) as

\[ \text{LM} = \frac{1}{2} - \frac{NT}{(\tilde{u}' \tilde{u})^2} \left\{ \frac{1}{(T-1)} [\tilde{u}'(I_N \otimes \hat{\lambda}_T' - I_{NT}) \tilde{u}]^2 + \frac{1}{(N-1)} [\tilde{u}'(\hat{\lambda}_N' \otimes I_T - I_{NT}) \tilde{u}]^2 \right\} \] (22)

to be taken as \( \chi^2(2) \). Note that

\[ \tilde{u}' \tilde{u} = \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{u}_{it}^2 \]

\[ \tilde{u}'(I_N \otimes \hat{\lambda}_T') \tilde{u} = \sum_{i=1}^{N} \left( \sum_{t=1}^{T} \tilde{u}_{it} \right)^2 \]

\[ \tilde{u}'(\hat{\lambda}_N' \otimes I_T) \tilde{u} = \sum_{t=1}^{T} \left( \sum_{i=1}^{N} \tilde{u}_{it} \right)^2 \]

so that computation of the test statistic from OLS' residuals would be quite straightforward. 7

If the time effects \( \lambda_t \) are assumed to be absent throughout, as is sometimes specified, the LM statistic for testing homogeneity across individuals, i.e. \( H_0: \sigma^2_{\tilde{u}} = 0 \), would be

---

7 In both (22) and (23) the factors \( N/(N-1) \) and \( T/(T-1) \) could be ignored without affecting the asymptotic properties of the tests as \( N \to \infty \) and \( T \to \infty \).
3.6 Testing Diagonality of the Covariance Matrix in SUR Systems

Sometimes it is of interest to test the proposition that the contemporaneous covariance matrix of a system of seemingly unrelated regressions (SUR) is diagonal [e.g., in Albon and Valentine (1977)]. Writing the \( i \)'th equation as

\[
y_i = X_i \beta_i + \epsilon_i
\]

where \( y_i \) is a T-vector and \( X_i \) is a \( T \times K_i \) matrix for \( i = 1, \ldots, M \) equations, the full system can be represented in extended vector form as

\[
y = X\beta + \epsilon
\]

where \( y \) is an n-vector and \( X \) is \( n \times \sum K_i \), with \( n = MT \). The usual assumption is that the regression equations are related through contemporaneously correlated disturbances, i.e., \( \text{E}(\epsilon_i \epsilon_j') = \sigma_{ij} I_T \) for \( i, j = 1, \ldots, M \), giving \( \text{E}(\epsilon \epsilon') = \Omega = (\Sigma \otimes I_T) \) with \( \Sigma = \{\sigma_{ij}\} \). To test the hypothesis that the equations are completely unrelated (i.e., \( \sigma_{ij} = 0 \) for all \( i \neq j \) giving \( \Sigma = \Sigma_0 \) as a diagonal matrix), the LM procedure would seem to be a useful approach as it would require only OLS estimation which amounts to constrained maximum likelihood under a normality assumption.\(^8\) Note that this situation differs from the previous one in

\[
LM = \frac{1}{2} \frac{NT}{(\tilde{u}'\tilde{u})^2(T-1)} \left[ \tilde{u}'(I_N \otimes I_T') - (I_N)\tilde{u} \right]^2
\]

\[
= \frac{1}{2} \frac{NT}{(T-1)} \left\{ \left[ \sum_{i=1}^{N} \left( \sum_{t=1}^{T} \tilde{u}_{it}^2 \right) \right] - \left[ \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{u}_{it}^2 \right] \right\}^2
\]

(23)

\(^8\) Exactly the same coefficient estimates will be given by application of OLS either to each equation (24) separately or to the \( n = MT \) observations in (25) as one regression.
that the covariance matrix of the disturbances will be diagonal but not necessarily scalar under the null hypothesis. Thus \( \Sigma_o = (\Sigma_o \otimes I_T) \) will not generally have the form \( \sigma^2 I \) although \( \Sigma_o = \text{diag}\{\sigma_{11}, \ldots, \sigma_{MM}\} \) still permits efficient estimation by OLS in this case.

Denote by \( \theta \) the vector of \( \frac{1}{2} M(M+1) \) distinct elements of \( \Sigma \) and, following Richard (1975), define the \( \frac{1}{2} M(M+1) \times M^2 \) selector matrices \( P \) and \( S \) such that \( \theta = S(\text{vec} \Sigma) \) and \( \text{vec} \Sigma = P'\theta \). These matrices have the properties that \( P'S = S'P \) and that \( \text{vec} \Sigma = P'S(\text{vec} B) \) where \( B \) is any symmetric \( M \times M \) matrix. Using these selectors, the general forms of the score vector and information matrix components corresponding to \( \theta \) parameters given as (16) and (17) above can be specialized to

\[
\begin{align*}
\delta_\theta(\theta, \beta) &= \frac{1}{2} P (\Sigma^{-1} \otimes \Sigma^{-1}) \text{vec}(E'E - TE) \\
I_{\theta\theta} &= \frac{1}{2} T P (\Sigma^{-1} \otimes \Sigma^{-1}) P'
\end{align*}
\]

where \( E = (e_1, \ldots, e_M) \) is the \( T \times M \) matrix such that \( \varepsilon = \text{vec} E \). Also, as shown by Richard (1975),

\[
I_{\theta\theta}^{-1} = 2T^{-1} S(\Sigma \otimes \Sigma) S'.
\]

Under the null hypothesis, the maximum likelihood estimate of \( \Sigma \) will be \( \tilde{\Sigma} = \tilde{\Sigma}_o = \text{diag}\{\tilde{\sigma}_{11}, \ldots, \tilde{\sigma}_{MM}\} \) with \( \tilde{\sigma}_{ii} = T^{-1} \tilde{e}_i' \tilde{e}_i \) where \( \tilde{e}_i \) is the vector of OLS residuals from the \( i \)th equation. Then the diagonal elements of \( (\tilde{E}'\tilde{E} - T\tilde{E}) \) will be zeros and the off-diagonal elements will be of the form \( \{\tilde{e}_i' \tilde{e}_j\} \). From the usual formula, using \( I_{\theta\theta}^{-1} = I_{\theta\theta}^{-1} \) and the properties of matrices \( P \) and \( S \),

\[
\begin{align*}
\text{LM} &= [\delta_\theta(\tilde{\theta}, \tilde{\beta})]' \tilde{I}_{\theta\theta}^{-1} [\delta_\theta(\tilde{\theta}, \tilde{\beta})] \\
&= \frac{1}{2} T^{-1} [\text{vec}(\tilde{E}'\tilde{E} - T\tilde{E})]' (\tilde{\Sigma}^{-1} \otimes \tilde{\Sigma}^{-1}) [\text{vec}(\tilde{E}'\tilde{E} - T\tilde{E})] \\
&= \frac{1}{2} T^{-1} \text{tr}[\tilde{\Sigma}^{-1/2}(\tilde{E}'\tilde{E} - T\tilde{E}) \tilde{\Sigma}^{-1/2}(\tilde{E}'\tilde{E} - T\tilde{E}) \tilde{\Sigma}^{-1/2}] \\
&= \frac{1}{2} T \text{tr}(R^2)
\end{align*}
\]
where \( R = \{r_{ij}\} \) is a symmetric \( M \times M \) matrix with

\[
    r_{ij} = \begin{cases} 
    \frac{\hat{e}'_i \hat{e}_j}{(\hat{e}'_i \hat{e}_i \cdot \hat{e}'_j \hat{e}_j)^{1/2}} & \text{for } i \neq j \\
    0 & \text{for } i = j.
    \end{cases}
\]

Therefore, an alternative expression is

\[
LM = \sum_{i=1}^{M} \sum_{j=1}^{i-1} r_{ij}^2
\]  

(26)

and this quantity would be asymptotically distributed as \( \chi^2 \) with \( M(M-1) \) degrees of freedom, this being the number of independent restrictions involved in constraining the off-diagonal elements of \( \Sigma \) to be zeros.

The criterion given by applying the LM procedure is intuitively suggestive: after fitting by OLS, compute all of the distinct across-equation correlations between the OLS residuals and use the sum of their squares to perform a joint test using (26) above.

3.7 The Chow Test

The final example to be considered in this chapter demonstrates rather nicely the usefulness of the LM procedure as a unifying principle for tests that have been developed by other methods.

Chow (1960) examined the problem of testing equality of sets of regression coefficients when two separate subsamples are available. For a total of \( n = (n_1 + n_2) \) observations, the model is

\[
\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}
\]

or, more compactly,
\[ y = X\beta + \varepsilon \tag{27} \]

In the partitioned form, \( y_1 \) is an \( n_1 \)-vector and \( y_2 \) is an \( n_2 \)-vector, \( \beta_1 \) and \( \beta_2 \) each contain \( k \) unknown coefficients so that \( \beta \) in (27) is a 2\( k \)-vector, and it is assumed that \( \varepsilon \sim N(0, \sigma^2 I_n) \). The null hypothesis imposes \( k \) restrictions on \( \beta \) as \( \beta_1 = \beta_2(=\hat{\beta}_2) \) to give the constrained model as

\[
y = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta^* + \varepsilon \equiv Z\beta^* + \varepsilon . \tag{28}\]

Under the null hypothesis, constrained maximum likelihood estimates are given by applying OLS to (28); these are denoted by \( \tilde{\beta}^* \) giving residuals \( \tilde{\varepsilon} = y - Z\tilde{\beta}^* \) and variance estimate \( \tilde{\sigma}^2 = n^{-1} \tilde{\varepsilon}'\tilde{\varepsilon} \).

If \( \text{rank}(X_1) = k < n_1 \) and \( \text{rank}(X_2) = k < n_2 \), the problem is simply one of testing linear restrictions on the coefficients of the linear regression (27). Denoting the unrestricted OLS estimates by \( \hat{\beta}' = (\hat{\beta}_1', \hat{\beta}_2') \) giving residuals \( \hat{\varepsilon}' = (\hat{\varepsilon}_1', \hat{\varepsilon}_2') \), the LM statistic is given from §2.4 (or from (10) above) as

\[
LM = n(\hat{\varepsilon}'\hat{\varepsilon} - \tilde{\varepsilon}'\tilde{\varepsilon})/\tilde{\varepsilon}'\tilde{\varepsilon}
\]

to be taken as \( \chi^2(k) \) from the asymptotic distribution. Conditional upon exogenous regressors, the transformation

\[
\frac{(n-2k)}{k} \frac{LM}{(n-LM)} = \frac{(\hat{\varepsilon}'\hat{\varepsilon} - \tilde{\varepsilon}'\tilde{\varepsilon})/k}{\tilde{\varepsilon}'\tilde{\varepsilon}/(n-2k)}
\]

is distributed as \( F(k, n-2k) \) when the hypothesis is correct. As expected, this is precisely Chow's criterion.

For the "updating" case, where \( \text{rank}(X_2) = n_2 < k \) and separate estimation of \( \beta_2 \) in the second subsample of (27) is not possible, Chow (1960)
developed a test by focusing on the prediction errors when the estimates from the first subsample are used to predict $y_2$. While Fisher (1970) observed that the same result would be given by the usual formula, once it was noted that residuals $\hat{e}_2$ would have to be identically zero and the degrees of freedom were adjusted accordingly, a more direct motivation is provided by the LM procedure. \(^9\)

For the unrestricted model (27),

$$d_\beta(\beta, \sigma^2) = \sigma^{-2} X' e = \sigma^{-2} \begin{bmatrix} X_1' e_1 \\ X_2' e_2 \end{bmatrix}$$

$$I_\beta = \sigma^{-2} X' X = \sigma^{-2} \begin{bmatrix} X_1' X_1 & 0 \\ 0 & X_2' X_2 \end{bmatrix}$$

where $I_\beta$ is singular with rank of $(k+n_2) < 2k$. As formulated in §1.8, the modified LM statistic for use when the information matrix is singular will be given by

$$LM = [d_\beta(\hat{\beta}, \hat{\sigma}^2)]' \tilde{I}_\beta [d_\beta(\hat{\beta}, \hat{\sigma}^2)]$$

where $\tilde{I}_\beta$ is (any) generalized inverse of $I_\beta$. The number of degrees of freedom of the asymptotic $\chi^2$ will be given by the number of independent overidentifying constraints, in this case by $n_2 = k-(k-n_2)$. A convenient approach is to use the Moore-Penrose g-inverse. \(^10\)

\(^9\) Zweifel (1976) proposed the LM procedure as an alternative to the Chow test when there are insufficient observations to estimate in a subsample, apparently without recognizing the connection.

\(^10\) It is not difficult to specify a vector $f$ such that $d_\beta(\hat{\beta}, \hat{\sigma}^2) = \tilde{I}_\beta f$, thereby demonstrating numerical invariance to the choice of g-inverse.
where \((X'_2X'_2)^+ = X'_2(X'_2X'_2)^{-2}X'_2\), and this gives the modified LM statistic as

\[
\text{LM} = \tilde{\sigma}^{-2} \left[ \tilde{\epsilon}'X'_1(X'_1X'_1)^{-1}X'_1\tilde{\epsilon} + \tilde{\epsilon}'_2\tilde{\epsilon}_2 \right]
\]

\[
= n(\tilde{\epsilon}'\tilde{\epsilon} - \tilde{\epsilon}'_1\tilde{\epsilon}_1)/\tilde{\epsilon}'\tilde{\epsilon}
\]

to be taken as \(\chi^2(n_2)\). Transforming to get the exact distribution,

\[
\frac{n_1-k}{n_2} \frac{\text{LM}}{(n-\text{LM})} = \frac{(\tilde{\epsilon}'\tilde{\epsilon} - \tilde{\epsilon}'_1\tilde{\epsilon}_1)/n_2}{\tilde{\epsilon}'_1\tilde{\epsilon}_1/(n_1-k)}
\]

is \(F(n_2, n_1-k)\) and this is precisely the criterion which was obtained by Chow from considering the prediction errors.

### 3.8 Further Aspects of the LM Test

One interesting feature of the LM test, particularly when it is used for misspecification testing in linear regression models, is that exact significance probabilities may sometimes be obtained when the corresponding W and LR statistics have distributions which are quite intractible. An example of such a situation was given in §3.4 where it was noted that the LM criterion for testing the addition of a nonlinearity to a linear model can be transformed to an F statistic when the explanatory variables are nonstochastic. Because the LM criterion is typically a function of the OLS residuals which are linear combinations of the disturbances, there are other situations in which exact distributions can be obtained fairly readily under a normality assumption. For example, Koerts and Abrahamse (1969) employed the Imhof (1961) method of numerical inversion.
of the characteristic function to obtain exact probabilities for the Durbin-Watson statistic which can be written as a ratio of quadratic forms in the OLS residuals. A similar approach could be used for the individual-effects-only version of the error components model, for which the LM statistic given as (23) above can be written as

$$\text{LM} = \left(\tilde{u}'Au/\tilde{u}'\tilde{u}\right)^2$$

where

$$A = \left[\frac{NT}{2(T-1)}\right]^{1/2} (I_N \otimes \hat{\epsilon}_T\hat{\epsilon}_T' - I_N)$$

and $\tilde{u}$ is the vector of OLS residuals. This method is examined in more detail in §6.5, in relation to the LM test for a class of heteroscedastic specifications.

When it is used as a framework for developing diagnostic tests, the LM principle requires an alternative parametric model to be specified with the current formulation being given by imposing restrictions upon the parameters of the full model. This need to specify a particular parametric alternative may seem to be unduly restrictive, especially as the alternative hypothesis is vague in many traditional misspecification tests. There appears to be a trade-off involved here: if the null hypothesis is rejected in a test with a particular alternative then a course of remedial action is clear, but a less specific diagnostic procedure might be expected to be more robust to incorrect formulation of the alternative hypothesis. One feature of the LM test, that will be observed in applications to testing for autocorrelation and heteroscedasticity in Chapters 4 and 6, is that the same criterion is sometimes given for a whole class of broadly similar alternatives. As a practical matter in testing for misspecification, the relative vagueness of the alternative hypothesis specification may be a desirable feature of the LM test. The econometrician is more likely to
have some general idea of the feared misspecification (e.g. autocorrelation of some order or heteroscedasticity related to some particular influence), than specific knowledge of the functional form of the alternative generalization.

The LM statistic appears in a variety of guises as criteria which are familiar to the econometrician but which have been motivated by other considerations. Some of these have been noted above and other connections between the LM principle and familiar criteria are examined in the next chapter. Engle (1978) has related Kmenta's (1967) test for a Cobb-Douglas production function and the tests for measurement error devised by Wu (1973) and Hausman (1978) to the LM procedure. The LM principle is implicit in Atkinson's (1970) presentation of the Cox (1962) procedure for discriminating between separate families of hypotheses. [See also Breusch and Pagan (1979b).]

Apart from the conceptual benefits as a unifying principle for misspecification testing, the LM test provides a convenient framework in which existing procedures can be extended and new tests can be developed. Advantages include asymptotic optimality under appropriate conditions, relative computational simplicity and, frequently, tractability of the exact finite sample distribution.
CHAPTER 4
TESTING FOR AUTOCORRELATION IN DYNAMIC LINEAR MODELS*

4.1 Introduction

If the disturbances are autocorrelated in a linear model with purely exogenous regressors, OLS estimates of the coefficient parameters will be consistent but inefficient under the usual assumptions. However when lagged values of the dependent variable are used as explanatory variables to give a dynamic formulation, OLS estimates will generally be inconsistent if autocorrelation is present.¹ For this reason, it is particularly important with dynamic models to have tests available for the misspecification error that is committed when autocorrelation is ignored and the model is estimated by OLS under the false assumption that the disturbances are serially uncorrelated.

In testing for autocorrelation as a misspecification error, full estimation of an alternative generalization requires the solution of non-linear normal equations. It is therefore considerably more difficult than OLS which would be the appropriate estimation technique if the null hypothesis of no autocorrelation was correct. The standard tests based on OLS residuals, notably the one proposed by Durbin and Watson (1950) and its extensions to higher order schemes by Schmidt (1972) and Wallis (1972), are attractive because they avoid estimation of the model with an autocorrelation process incorporated explicitly as would be required for W or LR tests to be used. But the standard tests, which were

* This chapter is to be published in a slightly different form as Breusch (1978). Godfrey (1978b), (1978c) has independently provided a similar analysis of tests for autocorrelation in dynamic models using the LM test framework.

¹ These ideas are discussed in most textbooks, e.g. Theil (1971, Ch.8).
noted in the preceding chapter to be related to the LM procedure, assume explicitly that all of the regressors are exogenous and this excludes, in particular, their valid application when lags of the dependent variable are used as regressors. Even though the standard tests would be biased toward acceptance of the null hypothesis of no serial correlation, Durbin (1970) showed that it is possible to obtain valid tests against autoregressive disturbances in dynamic models using residuals from OLS fitting. When proper account is taken of the interaction between the dynamics of the regression and the dynamic nature of autoregressive disturbances, a modification to the usual criterion is obtained to give a test with attractive large-sample properties but without the computational difficulties of estimating the full model. Thus it might reasonably be suspected that Durbin's test is closely related to the LM test when the latter allows for lags of the dependent variable to be included in the regressor set.

In this chapter, the LM test is developed for a variety of autocorrelation patterns in the disturbances of a dynamic model. The general methodology used by Durbin (1970) is discussed in §4.2 in relation to the LM procedure; the two approaches are not exactly the same in general but are shown to be quite closely related. In §4.3, the LM statistic for testing against the alternative hypothesis of an autoregression in the disturbances is obtained and compared with the statistic given by Durbin's method. The test statistic for the case where the process generating autocorrelated disturbances is a moving average instead of an autoregression is derived in §4.4 and application of the LM approach to testing for a mixed autoregressive-moving average process is also considered. Some special cases are treated in §4.5. Apart from the Durbin "h-statistic" appropriate for first-order autocorrelation alternatives,
other simplifications of the general form of the criterion are given for special cases including the important one for quarterly economic data of a joint first- and fourth-order process.

The relationship between Durbin's test statistic and that proposed by Box and Pierce (1970) may not be immediately apparent but it is shown in §4.5 how the latter may be obtained from the LM statistic by additional approximations. In §4.6, various regression strategies are given for computing the LM statistic. These avoid the problem of an undefined h-statistic and are more feasible computationally for higher order autocorrelation processes than direct generalizations of the h-statistic.

4.2 Durbin's Methodology and the LM Test

The general framework used by Durbin (1970) corresponds to the one discussed in §3.2 as the usual situation in which tests of misspecification are formulated. This framework has the alternative generalization parameterized by \( \theta' = (\theta_1', \theta_2') \) with the null hypothesis of no misspecification expressed parametrically by constraining a subset of the parameters to have specific values, say \( \theta_1 = \theta_{10} \). For this case, the LM statistic was given as

\[
LM = [d_1(\theta_{10}, \tilde{\theta}_2)]'\tilde{I}^{11}[d_1(\theta_{10}, \tilde{\theta}_2)]
\]

where \( d_1(\theta_1, \theta_2) \) is the subvector of the score, \( d(\theta) \), and \( \tilde{I}^{11} \) is the submatrix of the inverse of the information matrix, \( I^{-1} \), corresponding to the \( \theta_1 \) parameters under test, while \( \tilde{\theta}_2 \) contains maximum likelihood estimates under the null hypothesis of the parameters of the more restrictive model.

Durbin poses the problem of testing the hypothesis \( \theta_1 = \theta_{10} \), without
resorting to joint estimation of \( \theta' = (\theta_1', \theta_2') \) in the full model and using either \( W \) or LR criteria, as a two-stage procedure. Suppose that the restricted maximum likelihood estimate of \( \theta_2 \) has been obtained by solving \( d_2(\theta_1, \theta_2) = 0 \) for \( \tilde{\theta}_2 \) at the first stage and then \( d_1(t, \tilde{\theta}_2) = 0 \) is solved for an "estimate" \( t \) of \( \theta_1 \). The question then is: can \( t \) be used in something like a \( W \) statistic to provide a valid test of the hypothesis \( \theta_1 = \theta_{10} \)? What Durbin calls the "naive test" is to take \( t \) as having the same null hypothesis distribution as the maximum likelihood estimator of \( \theta_1 \) would have when \( \theta_2 \) is known and the known value of \( \theta_2 \) is used in the second stage to estimate \( \theta_1 \). If the elements of \( \theta_2 \) are known they are not parameters of the model so the information matrix for the full model is simply \( I_{11} \) corresponding to \( \theta_1 \) and the \( W \) test principle would lead to the criterion

\[
N = (t - \theta_{10})' \tilde{I}_{11}(t - \theta_{10}) \tag{2}
\]

to be taken as \( \chi^2(p) \) if there are \( p \) parameters fixed by the hypothesis \( \theta_1 = \theta_{10} \).

When the alternative hypothesis is that the disturbances of a linear model follow an autoregression, the two-stage estimation procedure would amount to applying OLS to get \( \tilde{\theta}_2 \) as the estimates of regression coefficients and then estimating the parameters of the autoregression as \( t \) from the OLS residuals. In this application, treating \( \theta_2 \) as known is equivalent to treating OLS residuals as if they are observations on the true disturbances. The naive test then corresponds to the usual joint test of significance of the coefficient parameters of an autoregression in observable variables.

However, the statistic based on a quadratic form in \((t - \theta_{10})\) which
has a $\chi^2(p)$ limiting distribution under the null hypothesis, when $t$ is obtained in the two-part procedure after estimating $\theta_2$, is shown by Durbin to be

$$D = (t - \theta_{10})' \hat{I}_{11} \hat{I}_{11} (t - \theta_{10})$$

(3)

where $I_{11}$ is the submatrix of $I$ corresponding to $\theta_1$ parameters. Now the difference between the matrices for which consistent estimates are used in the quadratic forms (2) and (3) is

$$I_{11} \hat{I}_{11} - I_{11} = I_{11}(I_{11}^{-1} - I_{11}^{-1})I_{11} = I_{12}I_{22}I_{21}$$

(4)

using $I_{11}^{-1} = I_{11}^{-1} + I_{11}^{-1} I_{12}I_{22}I_{21}I_{11}^{-1}$ from the partitioned inversion formula. Since $I$ is positive definite, $I_{11}^{-1}$ and hence $I_{22}$ are also positive definite and the matrix on the right hand side of (4) will be nonnegative definite. Unless $I_{12} = I_{21} = 0$, the relationship in large samples between Durbin's statistic and the naive one will be $D \geq N$ with generally $D > N$ because there is no particular reason for the vector $(t - \theta_{10})$ in (2) and (3) to be related to the weighting matrices. The naive test based on $N$ will therefore tend to understate the true significance of a calculated value of $t$, while the $D$ statistic in (3) can be referred to the $\chi^2(p)$ distribution to give approximate significance probabilities from the asymptotic distribution. Moreover, Durbin shows that the test using criterion $D$ has the same asymptotic power characteristics as the $W$ test (and, by implication, the same power asymptotically as the corresponding LR and LM tests). In the particular application of testing for autocorrelation, it was noted in §3.4 that $I_{12} = I_{21} = 0$ when the regressors are purely exogenous but in the next section it will be seen that this is not generally so when lags of the dependent variable are used as regressors.
It does appear that $D$ in (3) will be related to the LM statistic (1) by more than just asymptotic equivalence because, unlike $W$ and LR statistics, both $D$ and LM use aspects of the estimation problem under the constraint of the null hypothesis and neither requires that the full alternative model be estimated. While $D$, LM, $W$ and LR are all asymptotically equivalent, the first two are more closely related, especially when the statistic $D$ is attractive for its computational simplicity.

Suppose that the log-likelihood $\ell(\theta_1, \theta_2)$ is quadratic in the $\theta_1$ parameters which are under test so that $d_1(\theta_1, \theta_2) = \left[ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_1} (\theta_1, \theta_2) \right]$ is conditionally linear in $\theta_1$ given $\theta_2$. Then $d_1(t, \tilde{\theta}_2) = 0$ will be easily solved for $t$ and the second derivatives $\left[ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_1} (\cdot, \theta_2) \right]$ will not depend on $\theta_1$, so the expansion

$$d_1(t, \tilde{\theta}_2) = 0 = d_1(\theta_{10}, \tilde{\theta}_2) + \left[ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_1} (\cdot, \theta_2) \right](t - \theta_{10}) \quad (5)$$

will be exact, giving

$$d_1(\theta_{10}, \tilde{\theta}_2) = -\left[ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_1} (\cdot, \theta_2) \right](t - \theta_{10}). \quad (6)$$

Comparison of the LM statistic (1) and the $D$ statistic (3) using the relationship in (6) shows that the two will be exactly the same in this case provided: (i) the same estimate $\tilde{I}_{11}$ is used in both statistics, and (ii) the estimate of $I_{11}$ in (3) is $\tilde{I}_{11} = -\left[ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_1} (\cdot, \tilde{\theta}_2) \right]$. This is a rather tenuous basis upon which to describe Durbin's test as being the same as the LM test, but in the application for which the $D$ statistic was derived the log-likelihood is conditionally quadratic in the parameters under test so the difference between the criteria used in the two tests merely involves the question of the particular estimates.
that might be used for certain components of the information matrix.

In general, the expansion in (5) will be only an asymptotic approximation so that \( D \) and \( LM \) will be different in finite samples but, of course, asymptotically equivalent. Without parameters \( \theta_1 \) entering the log-likelihood in such a way that \( d_1(t, \hat{\theta}_2) = 0 \) is linear in \( t \), Durbin's approach will not be very useful because iterative calculations will be required anyway to solve for \( t \). But when Durbin's test for misspecification is computationally convenient, it differs from the \( LM \) test only by possible differences in the choice of estimate to be used for the information matrix.

4.3 Autoregressive Disturbances

Consider a linear model with up to \( m \) lags of the dependent variable included with \( q \) other variables as the regressors and with an autoregressive process of order \( p \) generating the disturbance:

\[
y_t = \gamma_1 y_{t-1} + \ldots + \gamma_m y_{t-m} + \beta_1 x_{1t} + \ldots + \beta_q x_{qt} + u_t
\]

(7)

\[
u_t = \alpha_1 u_{t-1} + \ldots + \alpha_p u_{t-p} + \varepsilon_t
\]

(8)

with \( \varepsilon_t \) serially independent for \( t = 1, \ldots, n \). The hypothesis to be tested is that the disturbance \( u_t \) in (7) is not autocorrelated; this corresponds to imposing \( p \) constraints on the parameters of the general model as

\[
H_0: \alpha_1 = \ldots = \alpha_p = 0.
\]

For some purposes, a more convenient matrix representation of the model is

\[
y_t = Y_t \gamma + X_t \beta + u_t = Z_t \delta + u_t
\]

(9)

\[
u_t = U_t \alpha + \varepsilon_t
\]

(10)
where

\[ Y = (y_1, \ldots, y_m)' \quad \beta = (\beta_1, \ldots, \beta_q)' \quad \alpha = (\alpha_1, \ldots, \alpha_p)' \]

\[ Y_t = (y_{t-1}, \ldots, y_{t-m}) \quad X_t = (x_{1t}, \ldots, x_{qt}) \quad U_t = (u_{t-1}, \ldots, u_{t-p}) \]

with \( Z_t = (Y_t : X_t) \) and \( \delta' = (\gamma', \beta') \). Alternatively, all observations can be written compactly as

\[ y = Yy + X\beta + u = Z\delta + u \quad (11) \]

\[ u = U\alpha + \epsilon \quad (12) \]

where \( y_t, Y_t, X_t, Z_t \) and \( U_t \) would be the t'th rows of \( y, Y, X, Z \) and \( U \) respectively so that \( Z = (Y : X) \). The null hypothesis is then \( \alpha = 0 \) for the vector of coefficients in the autoregression. Sometimes it will be more useful to represent the model using polynomials in the lag operator \( L \) where \( L^j y_t = y_{t-j} \) etc. Let

\[ \gamma(L) = 1 - \sum_{j=1}^m \gamma_j L^j \quad \text{and} \quad \alpha(L) = 1 - \sum_{j=1}^p \alpha_j L^j \]

enabling (7) and (8) to be written equivalently as

\[ \gamma(L)y_t = X_t\beta + u_t \]

\[ \alpha(L)u_t = \epsilon_t \]

It will be assumed that \( \epsilon_t \) is \( \text{NID}(0, \sigma^2) \) and that \( x_{jt} \) is exogenous so that \( E(x_{jt}\epsilon_{t'}) = 0 \) for all \( j = 1, \ldots, q \) and \( t, t' = 1, \ldots, n \). Both of the polynomial equations \( \gamma(L) = 0 \) and \( \alpha(L) = 0 \) are assumed to have all of their roots outside the unit circle implying a stable dynamic model in \( y_t \) and a stationary autoregressive disturbance, an assumption
which allows the initial values \( y_0, y_{-1}, \ldots, y_{-m+1} \) and \( u_0, u_{-1}, \ldots, u_{-p+1} \) to be ignored without affecting asymptotic properties of estimators or test statistics. Implicitly these initial values are treated as known constants or the processes are considered to have an infinite past, whichever is more convenient and, in the same way, summations over different ranges of \( t \) will generally be ignored because differences which are \( O_p(n^{-1}) \) are asymptotically negligible. Under these assumptions, the approximate log-likelihood is

\[
\ell(\alpha, \gamma, \beta, \sigma^2) = -\frac{1}{2} n \log (2\pi) - \frac{1}{2} n \log \sigma^2 - \frac{1}{2} \sigma^{-2} e' e
\]

where \( e \) is to be considered as the function of the data and the unknown parameters which would be given by rearranging (11) and (12).

Under the null hypothesis \( \alpha = 0 \), constrained maximum likelihood estimation of the unknown parameters is straightforward because then \( e = u \) in (11) and (12) so that the estimates are given by applying OLS to (11). These coefficient estimates will be denoted by \( \tilde{\delta} = (Z'Z)^{-1}Z'y \) giving residuals \( \tilde{e} = \tilde{u} = y - Z\tilde{\delta} \) and the residual variance estimate is \( \tilde{\sigma}^2 = n^{-1} \tilde{u}' \tilde{u} \). Define \( \tilde{U} \) formed from the lags of \( \tilde{u} \) and \( r = (r_1, \ldots, r_p)' = \tilde{U}' \tilde{U}/\tilde{u}' \tilde{u} \) so that \( r_j = \sum_t \tilde{u}_t \tilde{u}_{t-j} / \sum_t \tilde{u}_t^2 \) is the j'th autocorrelation coefficient \( (j = 1, \ldots, p) \) of the OLS residuals.

Now the hypothesis to be tested does not involve the variance parameter \( \sigma^2 \) which also is unrelated to the other parameters in the model in that the information matrix is block diagonal between \( \sigma^2 \) and \( \theta' = (\alpha', \gamma', \beta') = (\alpha', \delta') \). [See Durbin (1970, p.418).] Thus derivatives with respect to \( \sigma^2 \) can be ignored in forming the statistic, and the log-likelihood depends on the \( s = (p+m+q) \) parameters in \( \theta \) through the sum of squares function \( \sigma^{-2} e' e \). The required components of the score
vector and information matrix will then be given as in a nonlinear regression by

\[ d(\theta) = \sigma^{-2} G' \varepsilon \]

(13)

\[ I = \sigma^{-2} E(G'G) \]

(14)

where \( G \) is the \( n \times s \) matrix, \( G = -(\partial \varepsilon / \partial \delta)' \). From (11) and (12), noting that \( U \) will depend on the parameters in \( \delta' = (\gamma', \beta') \),

\[ G' = - \left( \frac{\partial \varepsilon}{\partial \theta} \right) = - \begin{bmatrix} \frac{\partial \varepsilon}{\partial \alpha} \\ \frac{\partial \varepsilon}{\partial \delta} \end{bmatrix} = \begin{bmatrix} U' \\ Z' + \frac{\partial \vec{U}}{\partial \delta} (\alpha \otimes I_n) \end{bmatrix}. \]

Under the null hypothesis \( \alpha = 0 \), the second term in \( (\partial \varepsilon / \partial \delta) \) will be zero, so as a first step in evaluating the quantities required for the LM statistic at the constrained estimates, \( G \) can be taken as \( G_0 = (U : Z) \). This simplifies calculation of the information matrix which will be, when the null hypothesis is correct,

\[ I = \sigma^{-2} E(G_0'G_0) = \sigma^{-2} E \begin{bmatrix} U'U & U'Z \\ Z'U & Z'Z \end{bmatrix}. \]

(15)

To construct the LM statistic, the score vector (13) and the information matrix (15) have to be evaluated at the restricted (OLS) estimates, \( \hat{\delta}' = (0, \hat{\delta}') \) and \( \hat{\sigma}^2 \). From (13),

\[ d(\tilde{\delta}) = \tilde{\sigma}^{-2} \tilde{G}' \tilde{\varepsilon} = \tilde{\sigma}^{-2} \tilde{G}' \tilde{u} = \tilde{\sigma}^{-2} \begin{bmatrix} \tilde{U}' \tilde{u} \\ \tilde{Z}' \tilde{u} \end{bmatrix} = \begin{bmatrix} nr \\ 0 \end{bmatrix} \]

where \( \tilde{G} = G_0 = (\tilde{U} : Z) \), and \( \tilde{Z}' \tilde{u} = 0 \) as the first-order conditions which give OLS as restricted maximum likelihood estimates. For the information
matrix estimate, one possible approach would be to take the expectation \( E(G'_0G_0) \) and then to replace unknown parameters by estimates and another would be to use simply \( \tilde{G}'\tilde{G} \) as the estimate of \( E(G'G) \). Durbin however uses a mixture of these two methods and the same choice of estimate for \( I \) will be used here for direct comparability of the LM statistic with Durbin's result. Firstly note that \( \tilde{V} = \tilde{\sigma}^2(Z'Z)^{-1} \) would be the usual estimate of the covariance matrix of the coefficient estimates that would be computed in the application of OLS which constitutes constrained estimation under the null hypothesis, so this component of \( \tilde{I} \) is already available. For the remainder of the submatrices of \( I \) in (15), expectations are taken when \( a = 0 \),

\[
E(u_{t-j}u_{t-k}) = E(c_{t-j}c_{t-k}) = \begin{cases} \sigma^2 & \text{for } j = k \\ 0 & \text{for } j \neq k \end{cases}
\]

\[
E(u_{t-j}x_{kt}) = 0 \text{ for all } j, k
\]

\[
E(u_{t-j}y_{t-k}) = E(c_{t-j} \cdot [\gamma(L)]^{-1} c_{t-k}) = \begin{cases} \sigma^2\psi_{j-k} & \text{for } j \geq k \\ 0 & \text{for } j < k \end{cases}
\]

where \( \psi \)'s are coefficients in the inverse of the polynomial in the lag operator on \( y_t \), i.e.

\[
[\gamma(L)]^{-1} = \sum_{j=0}^{\infty} \psi_j L^j \text{ with } \psi_0 = 1.
\]

These give \( E(U'U) = n\sigma^2I_p \), \( E(U'X) = 0 \) and \( E(U'Y) = n\sigma^2H \) where \( H \) is the \( p \times m \) matrix formed from the first \( m \) columns of: \(^2\)

---

\(^2\) The matrix given by Durbin (1970, p.420) appears to be incorrect in that it has \( (p+1) \) rows of which the last row is not required.
Collecting together all of these various components of the estimate of the information matrix gives

\[
\tilde{I} = n \begin{bmatrix}
I_p & \tilde{H} & 0 \\
\tilde{H}' & (n \tilde{V})^{-1}
\end{bmatrix}
\]

where \(\tilde{H}\) is formed as \(H\) using estimates of the \(\psi\) coefficients.

In the notation of §3.2, \(\theta_1 = \alpha\) and \(\theta_2 = \delta\) to give

\[
d_1(\tilde{\theta}) = nr \\
\tilde{I}^{ll} = n^{-1}(I_p - n\tilde{V}_{ll}\tilde{H}')^{-1}
\]

where \(\tilde{V}_{ll}\) would be the top left \(m \times m\) block of \(\tilde{V}\), i.e. the estimated covariance matrix of the \(\tilde{\gamma}\) estimates from OLS. The LM statistic as in (1) would then be

\[
LM = nr' (I_p - n\tilde{V}_{ll}\tilde{H}')^{-1}r
\]

and the LM test would reject the null hypothesis of no autocorrelation in the disturbances if the statistic exceeded the appropriate upper significance point from the \(\chi^2(p)\) distribution.

Instead of basing the test on the score vector, which in this situation is \(nr\) with \(r\) the vector of \(p\) residual autocorrelations, Durbin's test would use the vector \(a\) of "estimates" of \(\alpha\) obtained by regressing
\[ \tilde{u}_t \] upon \( \tilde{u}_{t-1}, \ldots, \tilde{u}_{t-p}, \) i.e.

\[ a = (\tilde{U}'\tilde{U})^{-1}\tilde{U}'\tilde{u}. \]

Rearranging, using \( r = \tilde{U}'\tilde{u}/\tilde{u}'\tilde{u} \) and \( \tilde{\sigma}^2 = n^{-1}\tilde{u}'\tilde{u}, \)

\[ \tilde{U}'\tilde{u} = (\tilde{U}'\tilde{U})a \]

\[ nr = [\tilde{\sigma}^{-2} \tilde{U}'\tilde{U}]a \quad (18) \]

which is the manifestation of equation (6) in this particular application.

Taking \( \tilde{I}_{11} = nI_p \) from (16), the statistic proposed by Durbin is

\[ D = na'(I_p - n\tilde{V}'\tilde{I}_1\tilde{V})^{-1}a \quad (19) \]

where, as Durbin notes: "The first \( p \) sample serial correlations have the same asymptotic distribution [as the estimates \( a \) of the coefficients in the autoregression] since in the null case they are asymptotically equivalent to the elements of \( a \". [Durbin (1970, pp.420-421), phrase in parentheses added.] In this case where the likelihood is quadratic in the \( a \) parameters under test, the discussion at the end of §4.2 is relevant: while in both (17) and (19) the same estimate \( \tilde{I}_{11} \) has been used, the two statistics \( LM \) and \( D \) differ only in that the estimate of \( I_{11} \) used in (19) is implicitly \( \tilde{I}_{11} = nI_p \). The two statistics would be exactly the same if \( \tilde{I}_{11} \) were to be estimated by simply evaluating second derivatives with respect to \( a \) using the restricted estimates of the parameters, i.e. by \( \tilde{I}_{11} = [\tilde{\sigma}^{-2} \tilde{U}'\tilde{U}] \).

There is one situation where the two statistics would coincide: if \( p = 1 \) then \( a_1 = r_1 \), except possibly for end effects due to summations over different ranges of \( t \). In that case, \( \tilde{H} = (1, 0, \ldots, 0) \) and \( r = r_1 \) is a scalar so that the \( LM \) statistic from (17) becomes
LM = \frac{n\hat{r}^2_1}{1 - n\hat{V}(\hat{\gamma}_1)} \quad \text{(20)}

where \( \hat{V}(\hat{\gamma}_1) \) is the estimated variance of the OLS estimate of \( \gamma_1 \), the coefficient of \( y_{t-1} \). Taking the square root,

\[ h = \sqrt{LM} = r_1 \sqrt{n/[1 - n\hat{V}(\hat{\gamma}_1)]} \]

where \( r_1 = a_1 = \sum_t \hat{u}_t \hat{u}_{t-1}/\sum_t \hat{u}_t^2 \), gives the familiar "h-statistic".

The so-called naive test would use \( N = a'\tilde{I}_{11}a \) as the criterion to be taken as \( \chi^2(p) \) under the null hypothesis. By the two choices for \( \tilde{I}_{11} \), this statistic could be variously \( na'a' \), or \( a'\hat{\sigma}^{-2} \hat{u}'\hat{U}a = n^2 r'[\hat{\sigma}^{-2} \hat{u}'\hat{U}]^{-1} r \) from (18), or \( nr'r \). These would be valid if the information matrix were to be block diagonal between the \( \alpha \) parameters and the coefficients \( \delta' = (\gamma', \beta') \) in the regression, but with lags of the dependent variable as regressors, the block of the information matrix connecting \( \alpha \) with \( \gamma \) comes from \( E(U'Y) \neq 0 \). In §4.6 some special cases are discussed in which the naive test is asymptotically valid or in which the same criterion can be employed with a different distribution used to indicate approximate significance probabilities.

### 4.4 Moving Average Disturbances

The LM statistic considered in the previous section is appropriate for testing the null hypothesis of no autocorrelation in the disturbances against an alternative which specifies the disturbance to be generated by an autoregressive (AR) scheme as in (8). Another important specification which would generate autocorrelated disturbances is the moving average (MA) process

\[ u_t = \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \ldots + \alpha_p \varepsilon_{t-p} \quad \text{(21)} \]
where \( \varepsilon_t \) is serially independent for \( t = 1, \ldots, n \). Replacing (8) by (21) gives a different alternative generalization to the null hypothesis of no autocorrelation in the disturbance of the dynamic regression model (7).

The principle advantage of the Durbin or LM approaches to testing for AR disturbances in a dynamic model is that the relatively complicated estimation problem under the alternative hypothesis is avoided. Fitts (1973) employed Durbin's methodology to obtain a test against MA disturbances in a dynamic regression model but this approach does not appear to be very practicable. He considered the simple case with one lag of the dependent variable and one exogenous variable as the regressors with a first-order MA in the disturbance \( (p = m = q = 1) \) so that Durbin's method would be: assume \( \alpha_1 = 0 \) and estimate \( \gamma_1, \beta_1 \) and \( \sigma^2 \) by OLS and then estimate \( \alpha_1 \) using the OLS residuals in place of the unobservable true disturbances. The difficulty with this is that, unlike estimating an AR in the residuals, estimation of even a simple MA is not a one-step operation because it would require iterative methods to solve nonlinear normal equations. Durbin's test therefore offers little, if any, advantage over LR or W tests in that its computational complexity rivals that of estimating the full model under the alternative hypothesis. Because the log-likelihood is not quadratic in the parameters under test, the Durbin test does not coincide with the LM procedure and while the former does not really avoid the difficulty of estimation in the full model, the latter does, as will now be shown.

The general model to be considered is the dynamic regression equation (7) or (9) or (11) with the disturbance \( u_t \) formed by the moving average process in (21). Parallel to (10) and (12) respectively, the MA can be written as
\[ u_t = E_t \alpha + \epsilon_t \]

\[ u = E\alpha + \epsilon \]

where \( E_t = (e_{t-1}, \ldots, e_{t-p}) \) is the \( t \)'th row of \( E \). Again the null hypothesis is \( \alpha = 0 \) and similar assumptions are made, including \( \epsilon_t \sim \text{NID}(0, \sigma^2) \) and omission of explicit treatment of initial conditions, to give the relevant part of the log-likelihood as the sum of squares function \(-\sigma^{-2} \epsilon' \epsilon\). Now \( \epsilon = u - E\alpha \) will depend on \( \alpha \) parameters directly and also through \( E \), and on \( \delta' = (\gamma', \beta') \) parameters through both \( u \) and \( E \). Thus for \( \theta' = (\alpha', \gamma') \),

\[ d(\theta) = \sigma^{-2} G' \epsilon \]

\[ I = \sigma^{-2} E(G'G) \]

where now

\[ G' = - \left[ \begin{array}{c} \frac{\partial \epsilon}{\partial \alpha} \\ \frac{\partial \epsilon}{\partial \delta} \end{array} \right] = \left[ \begin{array}{c} E' + \frac{\partial \text{vec} E}{\partial \alpha} (\alpha \otimes I) \\ Z' + \frac{\partial \text{vec} E}{\partial \delta} (\alpha \otimes I) \end{array} \right] \]

Under the null hypothesis \( \alpha = 0 \), the second terms in both \( (\partial \epsilon/\partial \alpha) \) and \( (\partial \epsilon/\partial \delta) \) will be zero. With the view being that the score vector and information matrix are to be evaluated at the restricted estimates under the null hypothesis, matrix \( G \) can be taken as \( G_o = (E : Z) = (U : Z) \) since \( E = U \) when \( \alpha = 0 \).

Restricted maximum likelihood estimates of the parameters \( \delta \) and \( \sigma^2 \) which are unknown under the null hypothesis will again be given by OLS so that the score vector becomes
where $\tilde{G} = \tilde{G}_0 = (\tilde{U}: \tilde{Z})$ as before. Alternative estimates of the information matrix would be given by $\tilde{I} = \sigma^{-2} \tilde{G}' \tilde{G}$ or by evaluating $\sigma^{-2} E(G'_o G_o)$ under the null hypothesis at the restricted (OLS) estimates. But these quantities are exactly the same as with an alternative hypothesis of AR disturbances. Irrespective of whether the process generating autocorrelation is hypothesized to be an AR or an MA of the same order, the LM procedure would lead to the same test statistic. In particular, Durbin's $h$-statistic which was derived for an alternative hypothesis of a first-order AR would also be the LM statistic for testing against a first-order MA.

It is interesting to consider what form the statistic would take if the disturbances were hypothesized to follow a composite autoregressive-moving average (ARMA) process. For simplicity, only the first-order specification

$$u_t = \phi_1 u_{t-1} + \varepsilon_t + \alpha \varepsilon_{t-1}$$

is considered, with the null hypothesis specifying $\phi_1 = \alpha_1 = 0$ so that $u_t$ is serially uncorrelated. However the method used previously will break down in this case because the information matrix of the full model will be singular when the null hypothesis is imposed on the parameters. If either the AR part or the MA part of (22) were to be specified as the autocorrelation process, the same score vector and information matrix would be obtained when the parameters are set to their values under the null hypothesis. Allowing both to be present would then introduce duplicate rows and columns into the information matrix to make it
singular. Also the score vector would contain repeated elements with the nonzero subvector being \( n(r_1 r_1)' \). The appropriate treatment when the information matrix is singular is to take a generalized inverse as was discussed in §1.8. When singularity is due to duplicated rows and columns, using any generalized inverse in the quadratic form would correspond to deleting one of the repeated rows and columns and deleting the corresponding element in the score vector, then forming the statistic in the usual way. The resulting statistic would obviously be the same as that for a simple AR or an MA alternative hypothesis.

4.5 Some Special Cases

There is one situation, apart from the obvious one where no lags of the dependent variable are used as regressors, in which the naive test would be valid. If the only \( y_{t-j} \) lags of the dependent variable in the regressor set are those with \( j > p \) where \( p \) is the order of the autocorrelation process under the alternative hypothesis, then \( H = n^{-1} \sigma^{-2} E(U'Y) \) is a null matrix, so from (17) the LM statistic is \( nr'r \) and from (19) Durbin's statistic for an AR alternative would be \( na'a \).

Another sort of simplification is possible when only the first lag of the dependent variable appears as an explanatory variable, i.e. \( m=1 \) so that \( \psi_j = \gamma_1^j \) and \( H = (1, \psi_1, \ldots, \psi_{p-1})' = (1, \gamma_1, \ldots, \gamma_1^{p-1})' \). If interest centres on testing for an autocorrelation process with just one parameter, i.e. the null hypothesis is \( a_k = 0 \) where \( a_j = 0 \) for all \( j \neq k \) under both null and alternative hypotheses, then \( H = \psi_{k-1} = \gamma_1^{k-1} \).

---

3 An alternative view of this is given by writing (22) as \( (1-\phi_1 L) u_t = (1+\alpha_1 L) e_t \) and noting that when \( \phi_1 = \alpha_1 = 0 \) there is a root common to both AR and MA operators. One of the usual conditions for identifiability of an ARMA model is not satisfied and this problem manifests itself as a singular information matrix.
The LM statistic would then be \( nr_k/[1 - n \tilde{y}^2(k-1) \tilde{V}(\tilde{y})] \) where \( \tilde{y} \) and \( \tilde{V}(\tilde{y}) \) are respectively the estimate of the coefficient of \( y_{t-1} \) and its estimated variance from estimation under the null hypothesis by OLS. This form of the LM statistic taken as a \( \chi^2(1) \) gives a convenient test against seasonal autocorrelation (e.g., \( k = 4 \) in a quarterly model).

Of particular interest in many econometric applications using quarterly data would be a test against joint first- and fourth-order autocorrelation as suggested by Wallis (1972), when the regressor set includes only the first lag of the dependent variable (m=1). From (17), the LM statistic would be

\[
n[r_1 \ r_4][1 - n \begin{bmatrix} 1 & 1 \\ \psi_3 & r_4 \end{bmatrix} \tilde{V}(\tilde{y})(1) \begin{bmatrix} r_1 \\ r_4 \end{bmatrix} ]^{-1} = \frac{n[r_1 - r_4^2 - n\tilde{V}(\tilde{y})(r_1 \tilde{y}_4 - r_4)^2]{1 - n\tilde{V}(\tilde{y})(1+\tilde{y}_4^2)}}
\]

where \( r_j = \sum \tilde{u}_t \tilde{u}_{t-j}/\sum \tilde{u}_t \) for \( j = 1,4 \) and \( \psi_3 = \tilde{y}_4^3 \) with all other quantities defined as before. This statistic would be taken as a \( \chi^2(2) \) under the null hypothesis.

There is one other important class of models where substantial simplifications may be made, but only with additional approximations that may be reasonable in some situations. Box and Pierce (1970) consider a model with no exogenous regressors, i.e. the autoregression

\[
\gamma(L)y_t = u_t \tag{23}
\]

where \( \gamma(L) \) is the \( m \) degree polynomial in the lag operator \( L \) that was defined on p.86 above, and they find the approximate asymptotic distribution of the first \( p \) (assumed \( p > m \)) autocorrelations of the least squares residuals, denoted here by the vector \( r \). Under the
hypothesis that the disturbances $u_t$ are serially independent, the Box-Pierce test takes

$$Q = nr'r = n \sum_{j=1}^{p} r_j^2$$

and treats this as $\chi^2(p-m)$ random variable to give a diagnostic check or a "portmanteau" test on the adequacy of (23) as the fitted model. Now this criterion would be the naive LM statistic for testing the null hypothesis $a = 0$ where the $p$ elements in vector $a$ parameterize an alternative hypothesis generalization that the disturbance follows a linear process (AR or MA) of order $p$. From the previous discussion in §4.3, it is clear that (24) would be asymptotically distributed as $\chi^2(p)$ not $\chi^2(p-m)$ if all the regressors were exogenous, but in (23) the regressors are all lags of the dependent variable. It would be interesting to obtain the correct LM statistic for comparison with the Box-Pierce criterion.

In the manner of §4.3, the model (23) can be written as

$$y_t = Y_t \gamma + u_t$$

for $t = 1,\ldots,n$ or as

$$y = Y \gamma + u$$

where $Y_t = (y_{t-1},\ldots,y_{t-m})$ is the $t$'th row of $Y$ and $\gamma = (y_1,\ldots,y_m)'$. Defining $U_t = (u_{t-1},\ldots,u_{t-p})$ as the $t$'th row of $U$, the information matrix for $\theta' = (a',\gamma')$ under the null hypothesis $a = 0$ would be

$$I = \sigma^{-2} E \begin{bmatrix} U'U & U'Y \\ Y'U & Y'Y \end{bmatrix} = n \begin{bmatrix} I & H \\ H' & W \end{bmatrix}$$

to give the LM statistic as [cf. expression (17)]
nr'\left( I_p - \tilde{H} \tilde{W}^{-1} \tilde{H}' \right)^{-1} r.

(25)

The definitions of $H$ and $\tilde{H}$ were given in §4.3 but previously
$n\tilde{W} = \sigma^{-2} E(Y'Y)$ was estimated simply as the covariance matrix of the OLS
estimates of the $\gamma$ parameters, i.e. as $\tilde{W} = (n\tilde{V})^{-1}$ where in this case
$\tilde{V} = \tilde{\sigma}^2(Y'Y)^{-1}$. An alternative estimate could be made after first taking
the expectation under the null hypothesis. Letting $W = \{w_{jk}\}$ for
$j,k = 1,\ldots,m,$

$$w_{jk} = n^{-1} \sigma^{-2} E \left[ \sum_t y_{t-j} y_{t-k} \right]$$

$$= n^{-1} \sigma^{-2} E \left[ \sum_t \left[ \gamma(L) \right]^{-1} u_{t-j} \cdot \left[ \gamma(L) \right]^{-1} u_{t-k} \right]$$

$$= \sum_{L=0}^{\infty} \psi_L \psi_{|j-k|}$$

(26)

where, as before, $[\gamma(L)]^{-1} = \sum_{j=0}^{\infty} \psi_j L^j$ with $\psi_0 = 1$.

From the definition of the matrix $H$, the $(j,k)'th$ element of $H'H$
would be

$$\sum_{L=0}^{p-g} \psi_L \psi_{|j-k|}$$

(27)

where $g = \max(j,k)$. At least for large values of $(p-g)$, expression
(27) could be used as an approximation to (26) giving $W \approx H'H$ and an
alternative estimate of $W$ as $\tilde{W} = \tilde{H'}\tilde{H}$. But with this approximation,
the matrix to be inverted in forming the LM statistic (25) would be

$[I_p - \tilde{H}(H'H)^{-1}H']$ which is singular so that (25) is undefined. Viewing
this in terms of the true covariance matrix rather than its estimate,
$\sqrt{n} r$ is asymptotically normally distributed under the null hypothesis
with a mean vector of zeros and a covariance matrix which is approximately

$[I - H(H'H)^{-1}H']$. This matrix is idempotent of rank $(p-m)$ implying that
Q = nr'r will have an asymptotic distribution approximating a $\chi^2(p-m)$ when the null hypothesis is correct.

Comparing (26) with (27), it can be seen that obtaining the Box-Pierce Q statistic from the LM statistic involves truncation of terms in $\psi_L$ which are the coefficients in the expansion of the inverse of $\gamma(L)$. The worst approximation errors would occur for $j$ or $k$ equal to $m$ when the summation in (27) is truncated after $(p-m)$ terms. For this truncation error to be small it would be required that $p$ be much larger than $m$ or that the roots of $\gamma(L) = 0$ be well outside the unit circle so that the $\psi_L$ coefficients converge rapidly to zero as $L$ increases. If applicable, the Box-Pierce test is computationally very simple, but using $\chi^2(p-m)$ as the distribution of $Q = nr'r$ requires $p > m$ and usually $p >> m$. Also, the approximation to arrive at the Box-Pierce result would not be valid if exogenous regressors had been fitted to the maintained model. The proper LM test on the other hand has no such restrictions upon it and may be applied for $p \leq m$ and when the fitted model includes exogenous regressors as well as lags of the dependent variable.

This connection with the LM criterion sheds some light on what appears in Box and Pierce (1970) to be an arbitrary choice of the number of residual autocorrelations to be used in the test that they propose and under what conditions the test may have desirable properties. In effect, the Box-Pierce procedure tests the null hypothesis of no autocorrelation in the disturbances against an alternative that the disturbances follow a $p'$th order linear process which may be an AR, an MA or some composite ARMA process. If the Box-Pierce test closely approximates the LM test, then large-sample optimality properties for the former may be inferred from the latter, for an alternative hypothesis which is a $p'$th order
linear process. However in typical applications it is apparent that the investigator is not really entertaining an alternative hypothesis as general as that implicit in the Box-Pierce test because the usual procedure if the test rejects the null hypothesis is to seek an extension involving one or two parameters that satisfactorily explains the residual autocorrelation. Thus it is not surprising that an LR test on the additional parameters of a selected overfitted model will frequently reject the more restrictive formulation in favour of the generalization, while the Box-Pierce $Q$ statistic is insignificant. Part of the problem comes from the requirement $p \gg m$ for $\chi^2(p-m)$ to be a good approximation to the asymptotic distribution of $Q$ and part is due to the structure of the criterion which involves autocorrelations at all orders up to and including $p$.

In the next section, various computational procedures for obtaining the proper LM statistic without the Box-Pierce approximation are given. The LM test, which is also applicable when the regressor set of the fitted model includes exogenous variables, allows more realistic alternative hypotheses to be considered, including $p \leq m$ and processes which correspond to overfitting the maintained model by just one parameter.

### 4.6 Alternative Forms of the LM Statistic

In §4.3, the estimate of the information matrix that was used in forming the test statistics was something of a hybrid. Some of the elements of $I$ were estimated by first taking expectations and then

---

4 See, e.g., Prothero and Wallis (1976), particularly the discussion by Chatfield. The authors in their reply indicate that they feel the main problem with the $Q$ statistic is due to inadequacy of the $\chi^2(p-m)$ distribution as an indicator of significance levels in small samples.
replacing unknown parameters by estimates while others come from simply evaluating negatives of second derivatives at the OLS estimates of the parameters. This was done to give direct comparability between Durbin's procedure and the corresponding LM test. For some simple cases, this mixed approach to estimating the information matrix provides a convenient way of forming the test statistic using the usual computational output from OLS estimation under the null hypothesis. However the matrix estimated in this way might not be positive definite, allowing the computed value of the statistic to be possibly negative and hence meaningless as a $\chi^2$ random variable. With a first-order process as the alternative, the $h$-statistic (20) would be undefined if $n\tilde{V}(\hat{\gamma}) > 1$.

One way to avoid this difficulty would be to use the same method for estimating all elements of the information matrix.

Consider the LM statistic for the alternative hypothesis that the disturbance follows an autoregressive process of order $p$ as developed in §4.3 but with the information matrix estimated as in the Gauss-Newton approximation by

$$\tilde{I} = \tilde{\sigma}^{-2} \tilde{G}'\tilde{G}$$

where $\tilde{G} = (\tilde{U} : Z)$. Unless the columns of $\tilde{G}$ are linearly dependent, and there is no reason in general for them to be so, this estimate will be a positive definite matrix. Then, with $d(\tilde{\theta}) = \tilde{\sigma}^{-2} \tilde{G}'\tilde{u}$, the LM statistic would be

$$LM = \tilde{\sigma}^{-2} \tilde{u}'\tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{u}$$

$$= n(\tilde{u}'\tilde{u})^{-1}\tilde{u}'\tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{u}$$

$$= n R^2$$
where $R^2$ is the usual coefficient of determination in the regression of $\tilde{u}$ against $\tilde{U}$ and $Z$. Thus, after fitting the model under the null hypothesis by OLS, the LM statistic to test against a $p$'th order process in the disturbances can be formed in a second least-squares regression, this time with $\tilde{u}_t$ as the dependent variable and with $\tilde{u}_{t-1}, \ldots, \tilde{u}_{t-p}$ plus the original $Z_t$ explanatory variables as the regressors. One advantage of this approach is that it is more flexible, allowing different dynamic structures in $y_t$ with different patterns of autocorrelation as the alternative hypothesis to be tested, but using the same general procedure and avoiding explicit derivation of the matrix $H$ defined on p.89. For example, suppose the maintained model is

$$y_t = \gamma_2 y_{t-2} + \gamma_3 y_{t-3} + x_t^\prime \beta + u_t$$

which has been estimated by OLS and it is desired to test the hypothesis that

$$u_t = \alpha_1 u_{t-1} + \alpha_4 u_{t-4} + \varepsilon_t.$$  

Then the LM statistic could be formed as $n R^2$ in the regression of the OLS residuals, $\tilde{u}_t$, on $\tilde{u}_{t-1}, \tilde{u}_{t-4}, y_{t-2}, y_{t-3}$ and $x_t$.

Several variants of this regression strategy are made possible by using different aspects of the second regression but yield statistics with exactly the same asymptotic properties. Firstly, instead of taking $R^2$ which is the ratio of explained to total sums of squares, the ratio of explained to residual sums of squares could be used. This can be seen by denoting the vector of residuals from the second regression by $e$ and letting $s^2 = n^{-1} e' e$ so that, from the sum of squares decomposition,
\[ \tilde{u}'\tilde{u} = e'e + \tilde{u}'\tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{u} \]

and dividing through by \( n\hat{o}^2 \) gives

\[ 1 = s^2/\hat{o}^2 + n^{-1} \hat{o}^{-2} \tilde{u}'\tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{u} . \]  

(29)

Here the second term on the right hand side is \( n^{-1}LM \) with \( LM \) defined as in (28) so that \( s^2/\hat{o}^2 - 1 \) in probability when \( LM \) has a limiting distribution. Consequently the asymptotic properties of the statistic are unaffected if \( s^2 \) is used in place of \( \hat{o}^2 \) to give

\[ s^{-2} \tilde{u}'\tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{u} \]

(30)

which will typically be numerically larger than (28) because \( \tilde{u}'\tilde{u} \geq e'e. \)

Both (28) and (30) can be interpreted as statistics for testing the overall significance of the regression of \( \tilde{u} \) upon \( \tilde{G} = (\tilde{U} : Z) \).

Expression (30) which uses the variance estimate under the hypothesis that all coefficients are zero would correspond to the \( LM \) statistic while (30) which uses the residual variance estimate from the regression is analogous to a \( W \) statistic.\(^5\) Alternative forms of both (28) and (30) are given by noting that, since \( Z'\tilde{u} = 0, \)

\[ \tilde{u}'\tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{u} = \tilde{u}'\tilde{U}(\tilde{U}'\tilde{U} - \tilde{U}'Z(Z'Z)^{-1}Z'\tilde{U})^{-1}\tilde{U}'\tilde{u} \].

(31)

If a number of alternative hypotheses are to be tested, as is implicit in computer programs for time series analysis which calculate residual autocorrelations routinely for diagnostic checking purposes, the \( LM \) statistics can be obtained for each alternative hypothesis from

---

\(^5\) Of course, the assumption of fixed regressors which is required for transformation to the usual \( F \) statistic as in §2.4 is not justified in this situation.
expression (31) with some savings in computational effort. Typically 
\((Z'Z)^{-1}\) will be available within the program so the only additional
matrix to be inverted is of the same dimension as the number of auto-
correlations being jointly tested. For a test on the individual auto-
correlation \(r_j = \hat{u}_j'\hat{u}/\hat{u}'\hat{u}\), the LM statistic from (28) is
\[
\frac{\hat{u}'\hat{u}}{n} - \hat{u}_j'Z(Z'Z)^{-1}Z'\hat{u}_j^{-1}\hat{u}_j
\]
\[
= n r_j^2 (\hat{u}'\hat{u})^{-1} - \hat{u}_j'Z(Z'Z)^{-1}Z'\hat{u}_j^{-1}
\]
\[
\approx n r_j^2 [1 - n^{-1}\hat{u}'Z(Z'Z)^{-1}Z'\hat{u}_j^{-1}]
\]
where all quantities to be inverted except \((Z'Z)\) are scalars. Then
instead of using \(n r_j^2\) as approximately \(N(0,1)\) under the null
hypothesis, an estimate of the correct asymptotic standard error to be
attached to \(n r_j^2\) can be obtained quite readily as
\[
[1 - n^{-1}\hat{u}'Z(Z'Z)^{-1}Z'\hat{u}_j^{-1}]^{1/2}.
\]

In the regression of \(\hat{u}\) upon \(\hat{U}\) and \(Z\), the least-squares estimate
of the coefficients of \(\hat{U}\) would be
\[
d = [(\hat{u}'\hat{U} - \hat{u}'Z(Z'Z)^{-1}Z'\hat{u})^{-1}]n r_j^2
\]
so that (30) can also be written as
\[
s^{-2} d'[(\hat{u}'\hat{U} - \hat{u}'Z(Z'Z)^{-1}Z'\hat{u})^{-1}]d
\]
which is the quantity that would be used in a \(W\) test of the joint
significance of the coefficients of \(\hat{U}\). Durbin (1970) recommended this
approach as an alternative procedure to follow when the statistic \(D\)
using the mixed estimate of the information matrix is negative.

Another variant of the LM test would be given by regressing \(y\)
instead of \(\hat{u}\) upon \(\hat{U}\) and \(Z\) and testing the joint significance of
the coefficients of \( \tilde{U} \) (but not by testing the overall significance of this regression). Writing the estimated model under the null hypothesis as

\[
y = Z\delta + \tilde{u}
\]

or

\[
= \tilde{G} \begin{bmatrix} 0 \\ \tilde{\delta} \end{bmatrix} + \tilde{u}
\]

the estimated coefficients in the second regression with \( y \) as dependent variable would be

\[
(\tilde{G}'\tilde{G})^{-1}\tilde{G}'y = \begin{bmatrix} 0 \\ \tilde{\delta} \end{bmatrix} + (\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{u}.
\]

Thus the coefficients of \( \tilde{U} \) are estimated the same with \( y \) for the dependent variable as with \( \tilde{u} \) for the dependent variable. Also the residual variance estimate would be

\[
n^{-1}[y'y - y'\tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'y]
\]

\[
= n^{-1}[\tilde{u}'\tilde{u} - \tilde{u}'\tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{u}]
\]

\[
= s^2
\]

so that the estimated covariance matrix of all coefficients in the regression would be \( s^2(\tilde{G}'\tilde{G})^{-1} \). With the same coefficient estimates and the same estimated covariance matrix, the \( W \) statistic for testing the significance of the coefficients of \( \tilde{U} \) in the regression of \( y \) upon \( \tilde{U} \) and \( Z \) would be the same as that with \( \tilde{u} \) as the dependent variable.

All of these alternative regression methods are variants of the "pseudo Wald test" form of the LM procedure that was derived in §3.2 where it was described as a test of the change in parameter estimates after one round of a numerical algorithm starting with estimates.
constrained by the null hypothesis as the initial parameter values. Hatanaka (1974) proposed a "residual adjusted Aitken estimator" given by one round of the Gauss-Newton algorithm as an efficient two-step estimator for dynamic models with AR disturbances. For consistency and efficiency of the two-step estimator, consistent initial estimates are required and these could be obtained by, say, a first step using instrumental variables. The LM test for the presence of the autoregression in the disturbances is equivalent to using the W test formula with estimates from the Hatanaka two-step procedure but with first step estimates from OLS. While OLS as the first step does not permit desirable properties such as consistency and efficiency under the alternative hypothesis to be attached to estimates from the second step, the relationship with the LM test indicates that a valid test can be based on these estimates.
CHAPTER 5
TESTING FOR AUTOREGRESSIVE DISTURBANCES IN
DYNAMIC SIMULTANEOUS EQUATIONS SYSTEMS

5.1 Introduction

Estimation of the parameters of a simultaneous equations system, with both serially independent and autoregressive disturbances, has been the subject of considerable attention in the econometrics literature. The complicated statistical nature of the simultaneous equations model has lead many writers to propose numerical methods for computing maximum likelihood estimates and to devise simpler alternative estimation schemes akin to two-step estimators. [See Hendry (1976) for a unifying survey.] As with the single equation (non-simultaneous) case, consistency and efficiency of the two-step estimator in the presence of both lagged endogenous variables and autocorrelated disturbances requires special treatment. However, for the model that would typically be taken as the alternative hypothesis when testing for autocorrelation in dynamic systems, Hatanaka (1976) has provided several two-step procedures which give efficient estimates.

Once the problem of obtaining maximum likelihood or asymptotically equivalent estimates for the full model with autoregressive disturbances has been solved, testing for the autocorrelation by the W or LR procedures is straightforward in principle. But the computational requirements to implement these approaches are often seen as being prohibitive and so they are little used in practice. The LM test, which requires estimation only under the null hypothesis of serial independence, offers advantages in reducing the computational burden, even in relation to using the W test procedure with efficient two-step estimation of the
full model. Especially for routine diagnostic testing, the LM procedure is more attractive for practical reasons than the other testing approaches to which it is asymptotically equivalent.

A distinction is usually made in estimation of simultaneous equations systems between "full information" methods, which use all of the prior restrictions on the coefficients, and "limited information" approaches, which estimate an individual equation neglecting the overidentifying restrictions on the other equations in the system. In the alternative hypothesis generalization with autocorrelated disturbances, full information is usually taken as including a vector autoregressive process and this is the model that is considered in §5.2 through §5.4. The specification of the full model including first-order dynamics and a first-order autoregressive process in the disturbances is given in §5.2 and the LM criterion for testing that the autoregressive parameters are all zero is obtained. In §5.3, the test which was proposed by Guilkey (1975) as a direct extension of the two-stage estimation method of Durbin (1970) is critically examined; a comment is made upon the amendments given by Maritz (1978) and a further correction to the Guilkey criterion is suggested. Computational aspects are considered in §5.4 using Hatanaka's various two-step methods as the framework for indirect calculation of the LM statistic. The savings in computational effort of the LM approach over consistent and efficient estimation of the full alternative hypothesis specification are indicated.

Limited information can have different meanings in the alternative hypothesis generalization depending upon the assumed structure of the autoregressive process in the disturbances or, from another viewpoint, upon what implications of a vector autoregressive process are ignored in estimating a single equation from a system. [See Amemiya (1966) and
Hendry (1976, Sect. 6).] Rather than extending the LM approach to a strictly-defined limited information framework, testing for autocorrelation in the disturbances of an individual equation from a dynamic simultaneous equation system is considered in §5.5 for the situation where a general instrumental variables procedure has been used for estimation under the null hypothesis. The two tests devised by Godfrey (1976), (1978a) are compared using the method for asymptotic power comparisons that was discussed in §1.5.

5.2 The Model and the LM Test

Consider a simultaneous equations system determining q endogenous variables and including first-order lags of the endogenous variables:

\[ Y = Y_T + Y_{-1}T + ZB + U \]  

(1)

Here \( Y \) and \( Y_{-1} \) are \( n \times q \) matrices containing observations on current and lagged endogenous variables respectively and \( Z \) is an \( n \times s \) matrix containing exogenous observations that are assumed to give \( n^{-1}Z'Z \) a finite nonsingular limit. Coefficient matrices \( T_0 \) and \( T_1 \) are each \( q \times q \) and \( B \) is \( s \times q \), where \( (I_q - T_0) \) is nonsingular and \( T_0 \) has zeros for its diagonal elements to accord with the normalization that is adopted in (1). The stochastic specification of the \( n \times q \) matrix of disturbances \( U \) is discussed below.

A more compact expression for the system is given by combining the predetermined variables as \( W = (Y_{-1} : Z) \) and their coefficients as \( A' = (T_1' : B') \) to give

\[ Y = Y_T + WA + U \]

An even more compact form is
\[ Y = XD + U \]

where \( X = (Y : Y_{-1} : Z) = (Y : W) \) and \( D' = (\Gamma'_0 : \Gamma'_1 : B') = (\Gamma'_0 : A') \).

Certain prior restrictions which exclude some of the variables from appearing in each of the equations are incorporated into the coefficient matrices as zeros; these are assumed to be adequate to render the remaining unknown parameters identifiable. Using a selector matrix \( S \) composed of zeros and units [as in Dhrymes and Erlat (1974)] enables the unrestricted elements of \( D \) to be collected into a vector \( \delta \) such that \( \delta = S'(\text{vec} \, D) \). Then, with the prior zero restrictions from normalization and exclusion imposed upon the elements of \( D \),

\[
\text{vec}(XD) = (I \ 0 \ X)\text{vec} \, D = (I \ 0 \ X)SS'(\text{vec} \, D) = (I \ 0 \ X)S\delta
\]

so that the model can be written in extended vector form with explicit zeros as

\[ y = (I \ 0 \ X)S\delta + u \quad (2) \]

where \( y = \text{vec} \, Y \) and \( u = \text{vec} \, U \).

For the alternative hypothesis generalization with autocorrelated disturbances, the specification of the disturbance in (1) is taken to be the vector autoregression

\[ U = U_{-1}R + E \quad (3) \]

where \( U_{-1} \) is \( U \) with each element lagged once (initial conditions are ignored), the rows of \( E \) are \( \text{NID}(0, \Sigma) \) and parameter matrices \( \Sigma \) and \( R \) are unrestricted except that \( \Sigma \) is taken to be symmetric positive definite. The hypothesis to be tested is that the disturbance \( U \) is not autocorrelated, as given by the \( q^2 \) constraints in
Corresponding to the vector representation of the model (2), the autoregression (3) can be expressed alternatively as

\[ u = (R' \Theta I_n)u_{-1} + \epsilon \]

or as

\[ u = (I_q \Theta U_{-1})\rho + \epsilon \]

where \( \rho = \text{vec } R \), \( u_{-1} = \text{vec } U_{-1} \) and \( \epsilon = \text{vec } E \) so that \( \epsilon \sim N(0, \Sigma \Theta I_n) \).

Thus an equivalent expression for the null hypothesis is \( H_0: \rho = 0 \).

In the full model with the autoregressive disturbance specification (3), the reduced form with serially independent disturbance is

\[ y = W\Pi_o + W_{-1}\Pi_1 + E(I - \Gamma_o)^{-1} \]

where

\[ \Pi_o = \begin{bmatrix} \Gamma_1 + (I - \Gamma_o)R \\ B \end{bmatrix} (I - \Gamma_o)^{-1}, \quad \Pi_1 = - \begin{bmatrix} \Gamma_1 R \\ BR \end{bmatrix} (I - \Gamma_o)^{-1} . \]

Under the null hypothesis in which \( R = 0 \), however, the reduced form simplifies to

\[ y = W\Pi + V \]

where

\[ \Pi = \begin{bmatrix} \Gamma_1 \\ B \end{bmatrix} (I - \Gamma_o)^{-1} \]

and the rows of \( V = U(I - \Gamma_o)^{-1} \) are serially independent.

Many estimators have been proposed for efficient full information estimation of the model under the constraint of the null hypothesis, when...
the rows of \( U \) are assumed to be serially independent. Most of these estimators, and certainly all of the more popular ones, can be represented as \( \hat{\delta} \) in the linearized estimating equation

\[
[S'(\hat{\Sigma}^{-1} \otimes \hat{X}'X)S]\hat{\delta} = S'(\hat{\Sigma}^{-1} \otimes \hat{X}')y 
\]

(8)

where \( \hat{X} = (\hat{Y} : W) \) with \( \hat{Y} = WH = W(W'W)^{-1}W'Y \) from the reduced form (6), and with \( \hat{Y} \) as a prediction of \( \bar{Y} \) using an estimate of \( \Pi \). The various estimators for the structural coefficients given as \( \hat{\delta} \) in (8) can be distinguished by their choices of initial estimates \( \hat{\Sigma} \) and \( \hat{Y} \) that are used in the linearization and by whether (or which of) the initial estimates are updated in an iterations scheme. While a full taxonomy may be found in Hendry (1976, Part I), some of the more common systems estimators may be mentioned. Three-stage least squares (3SLS) of Zellner and Theil (1962) would use \( \hat{Y} = WH = W(W'W)^{-1}W'Y \) from unrestricted OLS estimation of (6) to give the additional relationship \( \hat{Y}'Y = \hat{Y}'\hat{Y} \) that enables (8) above to be interpreted as a GLS regression solution, and \( \hat{\Sigma} \) would be taken from initial consistent estimates of the structural coefficients (usually by 2SLS). The estimator FIVE of Brundy and Jorgenson (1971) would use consistent instrumental variables estimates of the structural coefficients in \( \Gamma_0, \Gamma_1 \) and \( B \) to obtain \( \hat{\Sigma} \) and to derive an estimate of \( \Pi \) using (7). Full information maximum likelihood (FIML) would require the estimate of \( \hat{\Sigma} \) and that of \( \Pi \) in \( \hat{Y} = WH \) to be iterated upon until mutual reconciliation between all estimates in \( \hat{Y}, \hat{\Sigma} \) and \( \hat{\delta} \) was achieved.

It was noted in §1.9 that the LM test has the same asymptotic

---

1 See Hendry (1976, Part I) and for a more explicit instrumental variables interpretation see also Hausman (1975).
properties whether full maximum likelihood estimation of the constrained model is performed or whether other estimates which have the same limiting distribution when the null hypothesis is correct are used instead. Thus any of the estimators described in the preceding paragraph can be used to evaluate the formula for the LM statistic.

To form the statistic for testing the null hypothesis that \( R = 0 \) (or \( \rho = 0 \)) in the full model with autoregressive disturbances, the components of the score vector and of the inverse of the information matrix corresponding to \( \rho = \text{vec} \ R \) are required. Assuming that the difference equation in the endogenous variables and the autoregression in the disturbances are stable, the (approximate) log-likelihood for the full model is given by

\[
\ell(\rho, \delta, \Sigma) = -\frac{1}{2} n q \log(2\pi) - \frac{1}{2} n \log |\Sigma| + n \log \|I - \Gamma_0\| - \frac{1}{2} \text{tr}(\Sigma^{-1} E'E) \tag{9}
\]

where \( \| \cdot \| \) means the absolute value of the determinant, and

\[
\text{tr}(\Sigma^{-1} E'E) = \epsilon'(\Sigma^{-1} \otimes I)\epsilon
\]

can be related to the parameters through either the matrix or vector formulations of the model.

The score vector is obtained directly as

\[
d_{\rho}(\rho, \delta, \Sigma) = \partial \ell / \partial \rho = - (\partial \epsilon / \partial \rho)(\Sigma^{-1} \otimes I)\epsilon
\]

\[
= (\Sigma^{-1} \otimes U'_{-1})\epsilon
\]

\[
= \text{vec}(U'_{-1} E \Sigma^{-1})
\]

where, from (4), \( \partial \epsilon / \partial \rho = -(I \otimes U_{-1}) \). Evaluating the score vector at the restricted estimates, noting that \( E = U \) under \( H_0 \), gives the vector upon which the LM test is based as
\[
d_\rho(0, \tilde{\delta}, \tilde{\Sigma}) = (\Sigma^{-1} \otimes \tilde{U}_1')\tilde{u} = \text{vec}(\tilde{U}_1' \tilde{U} \Sigma^{-1}) \tag{11}
\]

where \(\tilde{u} = y - (I \otimes X)\delta\) and where \(\tilde{U}\) and \(\tilde{U}_1\) are formed from the restricted residuals.

Also required for forming the LM criterion is the corresponding component of the inverse of the information matrix, which may be written in the usual notation as \(I^{\rho\rho}\). Because there is a nonzero information link between \(\Gamma_0\) and \(\Sigma\) parameters in the simultaneous equations model, the parameters in \(\Sigma\) cannot be ignored in forming the information matrix as was done in the single equation situation of the previous chapter. However, instead of deriving the full information matrix for all of the parameters in the model and then obtaining the required submatrix of the inverse by partitioned inversion, an equivalent result can be obtained by firstly concentrating \(\Sigma\) parameters from the log-likelihood. This operation reduces the parameter set to \(\theta' = (\rho', \delta')\), and the resulting information matrix which permits \(\theta\) to be treated as the full parameter set is given (in its limiting form as \(n \to \infty\)) as \(D_\phi\) in Hatanaka (1976, p.193).

Evaluating Hatanaka's matrix \(D_\phi\) at \(R = 0\) gives, for the reduced (and reordered) parameter set \(\theta' = (\rho', \delta')\),

\[
I = \begin{bmatrix}
I_{\rho\rho} & I_{\rho\delta} \\
I_{\delta\rho} & I_{\delta\delta}
\end{bmatrix}
\]

where

\[
I_{\rho\rho} = n \Sigma^{-1} \otimes \Sigma
\]

\[
I_{\delta\rho} = I_{\rho\delta}' = S'(\Sigma^{-1} \otimes \bar{X}'U_1')
\]

\[
I_{\delta\delta} = S'(\Sigma^{-1} \otimes \bar{X}'\bar{X})S.
\]

For practical usage, estimates of the unknown parameters in \(\Sigma, \bar{X}\) and \(U_1\)
are required and these may be taken from the results of constrained estimation to give

\[ \tilde{I} = \hat{P}'(\hat{\Sigma}^{-1} \otimes I)\hat{P} \]

where \( \hat{P} = [(I \otimes \hat{U}_{-1}) : (I \otimes \hat{X})S] \), so that the estimate of \( \Sigma \) in \( I_{\rho\rho} \) is taken as \( \hat{U}_{-1}'\hat{U}_{-1} \).

Then, with \( I_{\rho\rho} = (I_{\rho\rho} - I_{\rho\delta}I_{\delta\delta}^{-1}I_{\delta\rho})^{-1} \) and using the score vector from (11), the LM criterion for testing the hypothesis of no autocorrelation becomes

\[ LM = \tilde{u}'(\hat{\Sigma}^{-1} \otimes \hat{U}_{-1}) I_{\rho\rho} (\hat{\Sigma}^{-1} \otimes \hat{U}_{-1}) \tilde{u} \quad (12) \]

which would be taken as \( \chi^2(q^2) \) from the asymptotic distribution.

5.3 Comments on Previous Formulations

It is interesting in this application to relate the LM statistic with the one that would be given by the Durbin (1970) two-stage estimation procedure discussed in §4.2. Guilkey (1975) used Durbin's framework to derive a test for first-order autoregressive disturbances in a dynamic simultaneous equations system, but the criterion given by Guilkey has been subject to a number of corrections by Maritz (1978). The close relationship between Durbin's procedure and the LM test provides a setting in which the test of Guilkey and the subsequent amendments by Maritz can be examined.

To implement the Durbin test, the first stage is to obtain the constrained estimates which have been denoted here by \( \rho = 0, \delta \) and \( \tilde{\Sigma} \) with residuals \( \tilde{u} \). The second stage is to "estimate" the parameters \( \rho \) as \( r \) by solving \( d_\rho (r, \delta, \tilde{\Sigma}) = 0 \). From (10), \( r \) is defined by
\[ d_\rho (r, \hat{\delta}, \hat{z}) = (\hat{z}^{-1} \otimes \hat{U}'_{-1}) [\hat{u} - (I \otimes \hat{U}_{-1}) r] = 0 \]

giving the relationship between the score vector upon which the LM test is based and the estimates \( r = \text{vec}[(\hat{U}'_{-1} \hat{U}_{-1})^{-1} \hat{U}' \hat{U}] \) to be

\[ d_\rho (0, \hat{\delta}, \hat{z}) = (\hat{z}^{-1} \otimes \hat{U}'_{-1}) u = (\hat{z}^{-1} \otimes \hat{U}'_{-1} \hat{U}_{-1}) r. \quad (13) \]

Since the log-likelihood (9) is quadratic in the parameters \( \rho \) conditional on the other parameters in the model, the discussion in §4.2 relating Durbin's procedure and the LM test is relevant. Except possibly for the estimate of \( \hat{I}_{\rho \rho} \) that is used in forming the criteria, the Durbin and LM statistics can be expected to be the same; with \( \hat{I}_{\rho \rho} \) estimated as \( \hat{I}_{\rho \rho} = (\hat{z}^{-1} \otimes \hat{U}'_{-1} \hat{U}_{-1}) \), which amounts to evaluating negatives of second derivatives at the constrained estimates, the criteria given by the two testing methods will be exactly the same. [Compare (13) above with expression (6) of Chapter 4 given on p. 84 above.] Then, from (12) using (13), an alternative expression for the LM statistic in terms of the two-stage estimate \( r \) (i.e. the Durbin criterion) would be

\[ \text{LM} = r' (\hat{z}^{-1} \otimes \hat{U}'_{-1} \hat{U}_{-1}) \hat{I}^{\rho \rho} (\hat{z}^{-1} \otimes \hat{U}'_{-1} \hat{U}_{-1}) r \]
\[ = r' \hat{I}_{\rho \rho} \hat{I}^{\rho \rho} \hat{I}_{\rho \rho} r. \quad (14) \]

As noted by Maritz (1978), the form of the criterion given by Guilkey (1975) is in error because it is (in the present notation)

\[ G = r' [\hat{I}^{\rho \rho}]^{-1} r \quad (15) \]

instead of (14). While the quantity \( G \) rather resembles a \( W \) statistic for testing the hypothesis \( \rho = 0 \), it cannot be a correctly formulated \( W \) statistic because that would require \( r \) to have the same null hypothesis limiting distribution as the estimate of \( \rho \) in full maximum likelihood estimation of the alternative hypothesis generalization.
Estimating $\rho$ as $r$ in the two-stage constrained approach will not give this property; if LM in (14) has the required $\chi^2(q^2)$ asymptotic distribution it will not generally be possible for (15) to have the same limiting distribution.

The difference between the matrices for which consistent estimates are used in (14) and (15) is

$$I_{\rho\rho} - \left[I^{\rho\rho}\right]^{-1} = I_{\rho\delta}(I_{\delta\delta} + I_{\delta\delta}^{-1})I_{\delta\rho}$$

which is nonnegative definite, so the formulation used by Guilkey would tend to understate the true significance of a calculated value of $r$ if it was taken as $\chi^2(q^2)$. In fact, in comparison with what Durbin calls the naive test using $N = r' I_{\rho\rho} r$, the difference between the weighting matrices in $G$ and $N$ is

$$I_{\rho\rho} - \left[I^{\rho\rho}\right]^{-1} = I_{\rho\delta} I_{\delta\delta}^{-1} I_{\delta\rho}$$

which is also nonnegative definite. Thus, the criterion formulated by Guilkey when used as a $\chi^2(q^2)$ would lead to worse understatement of the true significance levels than when the naive test is used.

Other problems with Guilkey's test relate to the expression that is given for the information matrix. In taking second derivatives of the log-likelihood, the parameters in the covariance matrix $\Sigma$ are ignored by invoking block-diagonality of the information matrix. Also, there appears to be no contribution in the second derivatives from the Jacobian term in the log-likelihood. However, replacing current endogenous variables in the expression that is given for the information matrix by their reduced form predictions will account for the Jacobian term and for the nonzero information link between covariance matrix parameters $\Sigma$ and the coefficients of the current endogenous variables in $\Gamma_0$. This
correction is noted briefly by Maritz.

A similar problem, but one which remains uncorrected by the Maritz amendments, comes from setting to zero the off-diagonal blocks of the information matrix connecting the autoregressive parameters $\Gamma$ with the coefficients in $\Gamma_0$. In Guilkey's derivations, current observations (endogenous and exogenous) are treated together, but while $\Gamma$ and $B$ are unrelated via the information matrix the same is not true of $\Gamma$ and $\Gamma_0$. Defining $\gamma_0 = \text{vec } \Gamma_0$ and recalling that $\rho = \text{vec } R$, it is not difficult to show that, even when $\Gamma = 0$,

$$- \frac{\partial^2 \ell}{\partial \gamma_0 \partial \rho'} = (\Sigma^{-1} \theta \gamma' U_{-1})$$

will generally have a nonzero expectation when the system is anywhere first-order dynamic. In a dynamic simultaneous system, current values of each endogenous variable will depend generally on the lagged values of all endogenous variables and hence they will be correlated with the disturbances at that lag, even when the disturbances are serially independent.

Rather than obtaining an explicit analytical expression for the information matrix, which would be more complicated still than expression (8) of Guilkey (1975, p.715) when all corrections have been made, this situation is one in which indirectly computing the LM criterion would be a useful approach.

5.4 Alternative Computational Methods

It was observed in §3.2 that the LM statistic can be computed indirectly using a procedure similar to efficient two-step estimation. But, while consistent and efficient estimation of the full alternative hypothesis generalization requires initial first round estimates to be
consistent for that specification, constrained efficient estimates are used in the indirect LM test. In the present application the initial estimates would be taken as \((0, \tilde{\delta}, \tilde{\Sigma})\) for parameters \((\rho, \delta, \Sigma)\), and the LM criterion would be given by the usual Wald test formula for the hypothesis that \(\rho = 0\), using the estimate of \(\rho\) from the second step. Each of the three alternative two-step estimation methods proposed by Hatanaka (1976) can be used for indirect computation of the LM statistic.

All three Hatanaka procedures share a common first step: obtain estimates of \(\rho, \delta\) and \(\Sigma\) by the method of instrumental variables in such a way that these estimates will be consistent for the full unrestricted model. Since the indirect LM test uses constrained estimates as initial values this step can be omitted. This would represent computational savings that could be substantial, especially in a misspecification testing situation where constrained estimates will already be available and where more than one alternative hypothesis may be of interest.

The three alternative second steps are either GLS regressions or generalized instrumental variables procedures and are distinguished by their choices of instruments and their choices of dependent variable. Defining

\[
P = [(I \otimes \hat{U}_{-1}) : (I \otimes X)S]
\]

and \(\hat{P}\) as in \(P\) but with \(\hat{X}\) replacing \(X\), the three Hatanaka alternatives using the constrained estimates as initial values would be as follows.

(a) Provided \(\hat{Y}\) in \(\hat{X} = (\hat{Y} : Y_{-1} : Z)\) is formed in unrestricted OLS estimation of the reduced form, \(^2\) run the GLS regression

\(^2\) The question of which reduced form permits this regression solution is discussed below.
(b) With \( \hat{Y} \) derived from the reduced form under \( H_0 \) (as in FIVE), run the generalized instrumental variables estimator

\[
\begin{bmatrix}
\hat{\rho}^* \\
\hat{\delta}^*
\end{bmatrix} = \left[ \tilde{P}'(\tilde{E}^{-1} \otimes I)\tilde{P} \right]^{-1} \tilde{P}'(\tilde{E}^{-1} \otimes I)y .
\]

(c) With \( \hat{Y} \) formed either as in (a) or (b), run the GLS regression

\[
\begin{bmatrix}
\hat{\rho}^{**} \\
\hat{\delta}^{**}
\end{bmatrix} = \left[ \tilde{P}'(\tilde{E}^{-1} \otimes I)\tilde{P} \right]^{-1} \tilde{P}'(\tilde{E}^{-1} \otimes I)y .
\]

In each case, the indirect LM statistic for testing the hypothesis will be given by the usual Wald test formula for testing the significance of the coefficients of \( \hat{U}_{-1} \). For example, method (c) gives the estimates as

\[
\begin{bmatrix}
\hat{\rho}^{**} \\
\hat{\delta}^{**}
\end{bmatrix} = \tilde{I}^{-1} \begin{bmatrix}
(\tilde{E}^{-1} \otimes \hat{U}'_{-1})\tilde{u} \\
0
\end{bmatrix}
\]

where \( S'(\tilde{E}^{-1} \otimes \hat{X}')\tilde{u} = 0 \) from constrained estimation under the null hypothesis. This gives

\[
\hat{\rho}^{**} = \tilde{I}^{\rho \delta} (\tilde{E}^{-1} \otimes \hat{U}'_{-1})\tilde{u}
\]

so that a Wald-type test of the significance of the coefficients of the lagged residuals would use the criterion

\[
\left( \hat{\rho}^{**} \right)' \left[ \tilde{I}^{\rho \delta} \right]^{-1} \left( \hat{\rho}^{**} \right) = \tilde{u}'(\tilde{E}^{-1} \otimes \hat{U}'_{-1})\tilde{I}^{\rho \delta} (\tilde{E}^{-1} \otimes \hat{U}'_{-1})\tilde{u}
\]

which is precisely the LM statistic as formulated in (12).
Method (b), which is a proper instrumental variables procedure, would produce a numerically different result but one which is asymptotically equivalent. Using

$$y = (I \otimes X)\delta + \tilde{u} = P \begin{bmatrix} 0 \\ \delta \end{bmatrix} + \tilde{u}$$

gives method (b) estimates as

$$\begin{bmatrix} \hat{\rho}^{**} \\ \hat{\delta}^{**} \end{bmatrix} = \begin{bmatrix} 0 \\ \delta \end{bmatrix} + [\hat{P}'(\check{\epsilon}^{-1} \otimes I)P]^{-1}\hat{P}'(\check{\epsilon}^{-1} \otimes I)\tilde{u}.$$ 

Then all that is required for $\hat{\rho}^{**}$ and $\hat{\rho}^{***}$ to give tests with the same asymptotic properties is

$$\operatorname{plim} [n^{-1} \hat{P}'(\check{\epsilon}^{-1} \otimes I)P] = \operatorname{plim} [n^{-1} \hat{P}'(\check{\epsilon}^{-1} \otimes I)\hat{P}]$$

under the null hypothesis and for local deviations from it, so the null hypothesis reduced form (6) is adequate for forming $\hat{Y}$.

What is not so clear, however, is that forming $\hat{Y}$ as the prediction in unrestricted OLS estimation of the null hypothesis reduced form (6) (as in 3SLS) is adequate for the instrumental variables estimator (b) to be run as a regression as in method (a). From Brundy and Jorgenson (1971, Th. 3, p.216) it follows that the regression solution (a) will reduce to the instrumental variables solution (b) when $\hat{Y} = W(W'W)^{-1}W'Y$ if and only if the columns of $\tilde{U}_{-1}$ are contained within the column space of $W$. In general then, methods (a) and (b) will produce numerically different results when $\hat{Y}$ is the same reduced form prediction that is used in estimating by 3SLS under the null hypothesis. But viewing method (a) as 3SLS estimation of an augmented model with $\tilde{U}_{-1}$ included in the structural equations as predetermined variables which are then truncated from the reduced form allows Theorem 5 of Brundy and Jorgenson
(1971, p.217) to be applied. Under the null hypothesis and for local deviations from it, consistent estimates of the null hypothesis reduced form (6) will also be consistent for the full reduced form (5), i.e. the reduced form coefficients of $\hat{U}_{-1}$ will be consistently estimated by zeros. Thus, truncation of the reduced form will allow the instrumental variables estimator to be run as a regression with unchanged limiting properties of the estimates.

These indirect solutions to computing the LM statistic or a close approximation to it provide ways of testing for autoregressive disturbances using standard computer programs. After estimation under the constraint of the null hypothesis of no autocorrelation, the model is re-estimated with $\hat{U}_{-1}$ included as predetermined variables which may be truncated from the reduced form. The usual joint test of the significance of the coefficients of the lagged residuals gives the indirect LM test.

In comparison with making inferences from consistent and efficient two-step estimation of the full model incorporating the disturbance process, the indirect LM test does not require a separate first step and the original instruments from estimation under the null hypothesis can be reused. Indirect LM testing methods also avoid explicit calculation of the information matrix as in Guilkey (1975) and Maritz (1978). Standard computer programs with facilities for reprocessing residuals from earlier runs can be used to calculate the LM statistic indirectly, or the additional calculations for autocorrelation diagnostics could be included in standard programs at small marginal computational cost. The indirect LM test generalizes readily to other situations including higher order (or no) dynamics in the model, higher order autoregressive processes as the alternative hypothesis and cases in which some of the $R$ matrix is specified a priori to be zero.
5.5 **Limited Information**

The LM approach could be applied to a limited information framework by paralleling the developments of §5.4, using Section 6 of Hatanaka (1976) where an efficient two-step limited information estimator is given for the dynamic simultaneous equations model with autoregressive disturbances. Rather than following this line, the relationship between the two tests devised by Godfrey (1976), (1978a) which use essentially arbitrary instrumental variables is examined.

Godfrey (1976) obtained a test for autocorrelation in the disturbance of a single equation from a dynamic simultaneous system by employing an extension of the method of Durbin (1970) that was discussed in §4.2. The test is based upon the residuals from the equation after constrained estimation under the null hypothesis, for which the method of instrumental variables has been used. Although some special choices of instruments are noted in that paper, they are taken to be essentially arbitrary provided the estimates are consistent when the null hypothesis of no autocorrelation is correct. An alternative test, proposed in Godfrey (1978a), uses the same residuals from the same estimates obtained by assuming that the null hypothesis is true, but differs in several respects from the one that was given earlier. This leads the proponent of the tests to remark that "(i)t would, therefore, be very interesting to have some evidence on the performance of these tests in finite samples" [Godfrey (1978a, p.227)]. Firstly however, it seems worthwhile to pursue asymptotic theory as far as practicable; in this section the concept of a local alternative hypothesis is used to compare the relative asymptotic power properties of the two tests.

The j'th equation from the dynamic simultaneous equations system may be written variously as
where all coefficient vectors include only unknown unrestricted parameters (hence the \( j \) subscript on the data matrices). To simplify notation the equation will be written without the \( j \) subscript as

\[
y = X\delta + u
\]  

(17)

so that all quantities in (17) have a different meaning from their previous usage in this chapter. The total number of included explanatory variables (current endogenous, lagged endogenous and exogenous) will be taken to be \( k \) so that in (17) \( y \) and \( u \) are \( n \)-vectors, \( X \) is an \( n \times k \) matrix and \( \delta \) is a \( k \)-vector of unknown coefficient parameters.

For the alternative hypothesis generalization, the disturbance in (17) is taken as following the autoregression

\[
u = \rho u_{-1} + \epsilon, \quad |\rho| < 1
\]  

(18)

where now \( \rho \) is a scalar and \( \epsilon \sim N(0, \sigma^2 I_n) \). The null hypothesis is

\[H_0: \rho = 0 \]

giving a serially uncorrelated disturbance in (17).

Instead of using maximum likelihood estimation, Godfrey adopts the autoregressive instrumental variables (AIV) method of Sargan (1959) in which estimates of the full model are given by minimizing the criterion function

\[
S(\delta, \rho) = [(y - \rho y_{-1}) - (X - \rho X_{-1})\delta]'Q[(y - \rho y_{-1}) - (X - \rho X_{-1})\delta]
\]
Here Q is the projector matrix

\[ Q = F(F'F)^{-1}F' \]

where F is an \( n \times g \) matrix of observations on \( g > k \) instrumental variables. Following Durbin (1970), a test of the hypothesis \( \rho = 0 \) is constructed by firstly minimizing \( S(\delta, 0) \) with respect to \( \delta \) to get the constrained estimates

\[ \hat{\delta} = (X'QX)^{-1}X'Qy \]
\[ = (\hat{X}'\hat{X})^{-1}\hat{X}'\hat{y} \]

where \( \hat{X} = QX \) so that \( \hat{X}'\hat{X} = \hat{X}'\hat{X} \), giving residuals \( \hat{u} = y - \hat{X}\hat{\delta} \) and variance estimate \( s^2 = n^{-1}\hat{u}'\hat{u} \). Then \( S(\hat{\delta}, \rho) \) is minimized with respect to \( \rho \) to obtain the "estimate"

\[ r = \frac{\hat{u}'_{-1}Q\hat{u}/\hat{u}'_{-1}Q\hat{u}}{\hat{u}'_{-1}u/\hat{u}'_{-1}\hat{u}} \]
\[ = \frac{\hat{u}'_{-1}\hat{u}/\hat{u}'_{-1}\hat{u}}{\hat{u}'_{-1}\hat{u}/\hat{u}'_{-1}\hat{u}} \]

where \( \hat{u} = Q\hat{u} \) and \( \hat{u}_{-1} = Q\hat{u}_{-1} \). Godfrey (1976) bases a test of \( H_0 \) upon the statistic

\[ \pi = \frac{n^{1/2}r}{s^2 n^{1/2} \left\{ (\hat{u}'_{-1}u_{-1} - \hat{u}'_{-1}X(\hat{X}'\hat{X})^{-1}X'\hat{u}_{-1}) / (\hat{u}'_{-1}u_{-1}) \right\}^{1/2}} \]

which is shown to be asymptotically distributed under \( H_0 \) as a standard normal deviate; alternatively \( \pi^2 \) may be taken as \( \chi^2(1) \).

The other testing criterion given by Godfrey (1978a) is based directly upon the autocorrelation coefficient (or autoregressive parameter estimate) obtained from the residuals, \textit{i.e.}

\[ \hat{r} = \hat{u}'_{-1}\hat{u}/\hat{u}'_{-1}\hat{u} \]

After showing that the usual formula for the h-statistic would usually
be invalid in this situation, a statistic using  \( r \) which would be \( N(0,1) \) when \( H_0 \) is true is given (in the present notation) as\(^3\)

\[
\theta = \frac{\sqrt{n}}{\sqrt{1 - 2\left(\hat{u}'_1 X(\hat{x}'\hat{x})^{-1}X'\hat{u}_1/n s^2\right) + \left(\hat{u}'_1 X(\hat{x}'\hat{x})^{-1}X'\hat{u}_1/n s^2\right)^2}} .
\]

For the present purpose, it will be convenient to simplify the notation for the two statistics by defining

\[
\Lambda = I_n - X(\hat{x}'\hat{x})^{-1}\hat{x}'
\]

which gives

\[
\hat{u} = y - X\hat{\delta} = Ay = \Lambda u
\]

since \( \hat{x}'\hat{x} = \hat{x}'X \), and also gives

\[
QA = F(F'F)^{-1}F' - \hat{x}(\hat{x}'\hat{x})^{-1}\hat{x}'
\]

as a symmetric, idempotent, \( n \times n \) matrix. Then the criterion \( \pi \) may be written more compactly as

\[
\pi = n^{-1/2} \hat{u}'_{-1} Q\hat{u}/\left[ s^2 n^{-1/2} \hat{u}'_{-1} QA\hat{u}_{-1} \right]^{1/2} \]  

(19)

and, using the asymptotically negligible approximation \( n^{-1/2} \hat{u}'_{-1} \hat{u}_1 \approx n^{-1} \hat{u}'_1 \hat{u}_1 \), a compact expression for the \( \theta \) statistic is

\[
\theta = n^{-1/2} \hat{u}'_{-1} \hat{u}_1/[ s^2 n^{-1/2} \hat{u}'_{-1} AA'\hat{u}_{-1} ]^{1/2} .
\]  

(20)

It may be helpful to record here that

\[
AA' = I_n - X(\hat{x}'\hat{x})^{-1}\hat{x}' - \hat{x}(\hat{x}'\hat{x})^{-1}X' + X(\hat{x}'\hat{x})^{-1}X' .
\]

Assumptions to be made are the same as those either given or implicit

\(^3\) The notation \( \theta \) for the statistic follows Godfrey (1978a) and is not to be confused with previous usage of \( \theta \) to denote parameters.
in Godfrey (1976) and (1978a); these assumptions and some of their immediate implications are collected together below.

(a) Each of the following second moment matrices is assumed to converge (in probability) to a finite limit:

\[ n^{-1} F'F \xrightarrow{p} M_1, \quad n^{-1} F'X \xrightarrow{p} M_2, \quad n^{-1} X'X \xrightarrow{p} M_3 \]

where \( M_1 \) is nonsingular \( g \times g \) and \( M_2 \) has full column rank of \( k \) so that \( (n^{-1} X'X)^{-1} \xrightarrow{p} (M_1M_2^{-1}M_1)^{-1} \).

(b) It is assumed that the instruments and lagged explanatory variables are contemporaneously uncorrelated in the limit with the innovation into the disturbance, and that

\[ n^{-1/2} F'\varepsilon \xrightarrow{p} N(0, \sigma^2 M_1) \]
\[ n^{-1/2} X'\varepsilon \xrightarrow{p} N(0, \sigma^2 M_3) \]

It is convenient to note at this stage that

\[ n^{-1/2} u'_{-1} \xrightarrow{p} N[0, \sigma^2 \lim (n^{-1} u'u)] \]

(c) Each of the following is assumed to have a finite probability limit:

\[ n^{-1} F'u_{-1}, \quad n^{-1} X'_{-1}u_{-1}, \quad n^{-1} X'u_{-1} \]

which will all be taken as nonzero for generality but some special cases are noted later.

(d) From assumptions in (a) and (c) above, it can be shown that each of the following has a finite probability limit:
\[ n^{-1} u_{-1} Q Au_{-1} \overset{P}{\to} v_1 \]
\[ n^{-1} u_{-1} Au_{-1} \overset{P}{\to} v_2 \]
\[ n^{-1} u_{-1} \Lambda \Lambda' u_{-1} \overset{P}{\to} v_3 \]

The analysis will be directed towards derivation of the noncentrality parameters of the \( \chi^2 \) limiting distributions of the statistics \( \pi^2 \) and, \( \theta^2 \), under the sequence of local alternative hypotheses

\[ H_A: \; \rho = n^{-\frac{1}{2}} \gamma \]

for some fixed \( \gamma \). This gives the disturbance process as

\[ u = \rho u_{-1} + \varepsilon = n^{-\frac{1}{2}} \gamma u_{-1} + \varepsilon \]

which implies, by the above assumptions:

\[ n^{-1} u'u \overset{P}{\to} \sigma^2 \]
\[ n^{-\frac{1}{2}} F'u \overset{P}{\to} N[\gamma \text{plim} (n^{-1} F'u_{-1}), \sigma^2 M_1] \]
\[ n^{-\frac{1}{2}} X'_1 u \overset{P}{\to} N[\gamma \text{plim} (n^{-1} X'_1 u_{-1}), \sigma^2 M_3] \]

Under the sequence of local alternatives, then, each of the following is bounded in probability:

\[ n^{-\frac{1}{2}} \hat{X}' u , \; n^{-\frac{1}{2}} X'_1 Q Au , \; n^{-\frac{1}{2}} X'_1 Au \]

and, in particular,

\[ n^{-\frac{1}{2}} u_{-1} Q Au \overset{D}{\to} N(\gamma v_1, \sigma^2 v_1) \]  \hspace{1cm} (21)
\[ n^{-\frac{1}{2}} u_{-1} Au \overset{D}{\to} N(\gamma v_2, \sigma^2 v_3) \]  \hspace{1cm} (22)

With these results to hand, the asymptotic distributions of the two
testing criteria $\pi$ and $\theta$ can be found for the local alternative hypothesis situation. Noting that $\tilde{u} = Au$ and $\tilde{u}_{-1} = u_{-1} - X_{-1}(\hat{X}'\hat{X})^{-1}\hat{X}'u$, the numerator of the statistic $\pi$ is given from (19) as

$$n^{-1/2} \tilde{u}_{-1}'Q\tilde{u} = n^{-1/2} u_{-1}'QAu - n^{-1/2} u'\hat{X}(\hat{X}'\hat{X})^{-1}X_{-1}'QAu$$

$$= n^{-1/2} u_{-1}'QAu + o_p(1)$$

as may be seen by writing the second term as

$$n^{-1/2}(n^{-1/2} \hat{X}'u)'(n^{-1} \hat{X}'\hat{X})^{-1}(n^{-1/2} X_{-1}'QAu).$$

The limiting distribution of the numerator of $\pi$ is therefore given by (21) above. From (20), the numerator of $\theta$ is

$$n^{-1/2} \tilde{u}_{-1}'\tilde{u} = n^{-1/2} u_{-1}'Au - n^{-1/2} u'\hat{X}(\hat{X}'\hat{X})^{-1}X_{-1}'Au$$

$$= n^{-1/2} u_{-1}'Au + o_p(1)$$

so that its limiting distribution is given by (22) above. Similar methods may be applied to the denominators of the two statistics to show that under the sequence of local alternatives

$$s^2 n^{-1} \tilde{u}_{-1}'QAu_{-1} \xrightarrow{p} \sigma^2 v_1$$

$$s^2 n^{-1} \tilde{u}_{-1}'AA'\tilde{u}_{-1} \xrightarrow{p} \sigma^2 v_3$$

which are the asymptotic variances of the respective numerators and correspond to the expressions obtained by Godfrey (1976), (1978a) when the null hypothesis is assumed to be true.

Therefore the asymptotic distributions under the sequence of local alternatives of the two statistics proposed by Godfrey have been found to be
\[ \pi^2 \overset{D}{\rightarrow} \chi^2[1, \text{nc}(\pi^2)] \text{, where } \text{nc}(\pi^2) = (\gamma^2/\sigma^2)\nu_1 \]

\[ \theta^2 \overset{D}{\rightarrow} \chi^2[1, \text{nc}(\theta^2)] \text{, where } \text{nc}(\theta^2) = (\gamma^2/\sigma^2)v_2^2/v_3 \cdot \]

A comparison of the asymptotic powers of the two tests involves the relative magnitudes of \( \nu_1 \) and the ratio \( v_2^2/v_3 \); whichever is larger indicates the more powerful test. Unfortunately it does not appear possible to obtain a relationship between the two noncentralities which holds for all regressor sets \( X \) and all possible sets of instruments \( F \) satisfying the conditions in (a), (b) and (c) above. But some special cases can be examined; in every case the test using \( \theta \) is found to be no less powerful, and in some cases more powerful, than the test based upon the criterion \( \pi \).

**Case I**

If the set of instruments \( F \) includes both \( y_{-1} \) and \( X_{-1} \), then

\[ Q\hat{u}_{-1} = Qy_{-1} - QX_{-1}\delta = y_{-1} - X_{-1}\delta = u_{-1} \]

so that

\[ u_{-1}'QAu_{-1} = u_{-1}'Au_{-1} = u_{-1}'AA'u_{-1} \]

which gives \( v_1 = v_2 = v_3 \), implying \( \text{nc}(\pi^2) = \text{nc}(\theta^2) \) and equal asymptotic power for the two tests. However, this outcome is somewhat obvious since with these instruments \( \hat{u}_{-1} = Q\hat{u}_{-1} \) giving \( r = \tilde{r} \) so the two procedures are testing the same quantity. In fact, with \( y_{-1} \) and \( X_{-1} \) included in \( F \) the two statistic defined by Godfrey will be numerically identical except for the approximation \( n^{-1}u_{-1}'u_{-1} \approx n^{-1}\tilde{u}'\tilde{u} \) that was made to obtain the expression for \( \theta \) in (20).
Case II

Suppose that only exogenous variables were to be used as instruments, i.e. $F$ does not contain any lagged endogenous variables. Then

$$\text{plim}(n^{-1}F'u_{-1}) = 0$$

implying

$$v_1 = 0, \quad v_2 = \sigma^2, \quad v_3 = \sigma^2 + \text{plim}(n^{-1}u_{-1}'X(\hat{X}'\hat{X})^{-1}X'u_{-1})$$

which gives

$$0 = \text{nc}(\pi^2) < \text{nc}(\theta^2)$$

Thus with this choice of instruments the $\pi$ test is asymptotically powerless while the $\theta$ test has some desirable properties including consistency. Godfrey (1978a, p.227) observes that the $\pi$ test cannot be used when the instruments are purely exogenous but the $\theta$ test has no such limitation. ⁴

Case III

Consider the situation where the explanatory variables $X$ are asymptotically uncorrelated with the lagged disturbance, i.e.

$$\text{plim}(n^{-1}X'u_{-1}) = 0.$$ 

This would give

$$v_1 = \text{plim}(n^{-1}u_{-1}'QAu_{-1}), \quad v_2 = \sigma^2, \quad v_3 = \sigma^2$$

where, from the idempotency of $QA$,

⁴ A further limitation on the use of the $\pi$ test noted by Godfrey is that $\pi$ requires $g > k$ instruments while $\theta$ requires only $g \geq k$. 
\[ u_{-1}' Q u_{-1} \leq u_{-1}' u_{-1} \]

implying \( v_1 \leq \sigma^2 \) and

\[ nc(\pi^2) \leq nc(\sigma^2) . \]

This situation would pertain, for example, if the system was not dynamic, and also in some cases of a recursive or partially recursive dynamic system. Thus with static, and possibly also with dynamic models, the \( \theta \) test is at least as powerful asymptotically as the \( \pi \) test and may be more powerful.

**Case IV**

The final special case to be considered is when

\[ \text{plim}(n^{-1} \hat{X}' u_{-1}) = \text{plim}(n^{-1} X' u_{-1}) \]

that is,

\[ \text{plim}[n^{-1}(X - \hat{X})' u_{-1}] = 0 . \]

In this case,

\[ v_1 = \text{plim}(n^{-1} u_{-1}' Q u_{-1}) - \text{plim}[n^{-1} u_{-1}' X(\hat{X}' \hat{X})^{-1} X' u_{-1}] \]
\[ v_2 = v_3 = \text{plim}(n^{-1} u_{-1}' u_{-1}) - \text{plim}[n^{-1} u_{-1}' X(\hat{X}' \hat{X})^{-1} X' u_{-1}] \]

so that from the idempotency of \( Q \), \( v_1 \leq v_2 = v_3 \) which gives

\[ nc(\pi^2) \leq nc(\sigma^2) . \]

One obvious situation in which this would occur is when the equation of interest includes no current endogenous variables as regressors and the explanatory variables in \( X \) are included in the instruments \( F \), for then \( \hat{X} = X \).

More generally, reverting to the notation with subscripts indicating
that the equation of interest is the $j$'th member of a set of simultaneous
equations, i.e. $X = X_j = (Y_j : W_j)$, consider the two-stage least squares
(2SLS) estimator which uses instruments $F = W$ with $W$ containing the
predetermined variables of the system. Then

$$\hat{X}_j = (\hat{Y}_j : W_j)$$

$$= [W(W'W)^{-1}W'Y_j : W_j]$$

so that

$$(X_j - \hat{X}_j) = [Y_j - W(W'W)^{-1}W'Y_j : 0]$$

$$= [V_j - W(W'W)^{-1}W'V_j : 0]$$

where $V_j$ is the submatrix of the reduced form disturbances corresponding
to the current endogenous variables that are included in the equation as
explanatory variables. Under local alternative hypotheses in which the
autoregressive parameter is $O(n^{-1})$, it is readily shown that

$$\text{plim}(n^{-1}W'V_j) = 0, \text{plim}(n^{-1}W'u_{-1}) \text{ is finite and } \text{plim}(n^{-1}V_j'u_{-1}) = 0$$

so that $\text{plim}[n^{-1}(X_j - \hat{X}_j)'u_{-1}] = 0$. Thus when 2SLS is used for estimation
under the null hypothesis of no autocorrelation, the $\theta$ test is again at
least as powerful as the $\pi$ test. 5

5.6 Concluding Comments

The indirect approaches to calculating the LM statistic in a full
information setting that are developed in §5.4 provide a straightforward
method of testing for autocorrelation in a dynamic simultaneous equations
system. Standard computer programs for estimation under the null
hypothesis of no autocorrelation can be reused to calculate the testing

5 I owe this point to a private communication with L.G. Godfrey.
criterion, or the statistic could be computed as a routine diagnostic in standard programs.

In the limited information setting, where the structure of the rest of the system would be ignored when testing for autocorrelation in the disturbance of an individual equation, two tests based upon the results of instrumental variables estimation under the null hypothesis have been compared for their relative asymptotic power. The \( t \) test given by applying Durbin's two-stage estimation procedure (which would coincide with the LM method using the same AIV criterion function) has been found to be no more powerful in many circumstances, and sometimes less powerful, than the \( \theta \) test based directly on the autocorrelation coefficient of the residuals from fitting the restricted model under the null hypothesis.
CHAPTER 6
TESTING FOR HETEROSCEDASTICITY INCLUDING RANDOM COEFFICIENTS*

6.1 Introduction

In many applications of the general linear model the usual assumptions, that the disturbance variance is constant and the regression coefficients are fixed parameters, may be questioned. Particularly with cross-sectional microeconomic data where observations may involve substantial differences in magnitude, it might be expected that the variance of the dependent variable as well as its mean can be related to some explanatory variable or variables. One way in which this may be characterized is to have regression coefficients varying randomly according to some probability distribution, either across subgroups of observations or across all observations. With time series of data too, heteroscedasticity may be present. If, for example, a relationship which satisfies the usual conditions for least-squares analysis when specified in real terms is instead estimated in nominal variables, the disturbance variance will be proportional to the square of the price level. When disturbances are heteroscedastic, the loss in efficiency incurred in misspecifying the disturbance and estimating by OLS may be substantial and, what is possibly more important, inferences from OLS estimates by the usual W tests may be seriously misleading because of biases in estimated standard errors.¹

The LM statistic for testing against a wide class of heteroscedastic specifications, including many of the models of random coefficient

* This chapter includes material to be published as Breusch and Pagan (1979a).

¹ Inefficiencies in estimating regression coefficients and biases in OLS standard error estimates in the presence of heteroscedasticity are discussed in most textbooks, e.g. Goldberger (1963, Sect. 5-4) and Johnston (1972, Sect. 7-3).
variation, is obtained in this chapter. As in most of the applications
of the LM test considered in previous chapters, the criterion in this
case can be computed in a least-squares regression, using the residuals
from OLS which would be the appropriate estimation technique if the
disturbances were homoscedastic. In this way, a test with the same
asymptotic properties in standard situations as the corresponding W or
LR tests is given without the iterative calculations that would be
required for maximum likelihood estimation of the full heteroscedastic
specification.2

The general model to be considered in this chapter is detailed in
§6.2 where its relationship to the many formulations of heteroscedasticity
that have been proposed is noted. Some of the difficulties associated
with full maximum likelihood estimation of regression models with certain
of these heteroscedastic specifications are mentioned briefly in §6.3.
In §6.4, the LM statistic is derived and arranged into a form that is
easily computed. Finite-sample properties are examined in §6.5, including
the moments of the statistic and numerical methods for obtaining exact
significance probabilities, and a bounds test procedure is used to analyze
a situation where it seems that the asymptotic $\chi^2$ approximation may be
inadequate. Further insights into the finite-sample properties of the
LM test are provided in Chapter 7 in which the results of a Monte Carlo
simulation study comparing it with W and LR tests are presented.

6.2 The Model with Heteroscedastic Disturbances

Consider the linear regression model with heteroscedastic disturbances:

$$y_t = x_t'\beta + u_t$$  \hspace{1cm} (1)

2 Estimation by maximum likelihood methods of various random
coefficient models is treated by Swamy (1970) and of other
heteroscedastic schemes by Goldfeld and Quandt (1972).
where $\beta$ is a $k$-vector of unknown coefficient parameters, $x_t$ is a $k$-vector of nonstochastic regressors, and

$$u_t \sim \text{NID}(0, \sigma_t^2) \quad \text{with} \quad \sigma_t^2 = h(z_t' \alpha)$$

(2)

for $t = 1, \ldots, n$. Function $h(\cdot)$, which is not indexed by $t$, is assumed to be continuous possessing at least first and second derivatives, and the argument of the function is a linear combination of exogenous variables which may or may not be related to the $x_t$ regressors. The first element of the $(p+1)$-vector $z_t$ is unity for all $t$ and $\alpha$ is a $(p+1)$-vector of unknown parameters functionally unrelated to the $\beta$ coefficients, thereby allowing the null hypothesis of homoscedasticity to be represented as $p$ restrictions on the parameters,

$$H_0: \alpha_2 = \ldots = \alpha_{p+1} = 0.$$  

(3)

Under the null hypothesis then, $z_t' \alpha = \alpha_1$ so that $\sigma_t^2 = h(\alpha_1) = \sigma^2$ is constant for all observations.

For estimation under the alternative hypothesis generalization, as would be required for $W$ or LR tests to be used, the functional form of $h(\cdot)$ is assumed to be known, although it will be seen that this knowledge is not required for the LM test to be performed. Also required would be the condition $h(z_t' \alpha) > 0$ which may be construed either as a restriction on the set of admissible $z_t$ values or as a constraint upon the parameter space of $\alpha$, given the observed $z_t$'s.

The representation in (2) is sufficiently general to cover most of the alternative hypotheses that have been distinguished in the literature. These are usually either

$$\sigma_t^2 = \exp(z_t' \alpha)$$

(4)

which was noted by Harvey (1976) to encompass the models of Park (1966),
Geary (1966) and Lancaster (1968), or

\[ \sigma_t^2 = (z_t^\alpha)^m \]  

(5)

with \( m \) a prespecified integer as in Rutemiller and Bowers (1968) with \( m = 2 \), Glejser (1969) with \( m = 1 \) or 2, Goldfeld and Quandt (1965), (1972) and Amemiya (1977) with \( m = 1 \). Formulation (4) includes the case where the variance is proportional to one of the regressors raised to some unknown power,

\[ \sigma_t^2 = \sigma^2 X_{lt}^\gamma \]

which is (4) with \( \alpha_1 = \log \sigma^2, \alpha_2 = \gamma \) and \( Z_{2t} = \log(X_{lt}) \). With \( m = 1 \), the formulation in (5) would also include many of the models of random coefficient variation that have been advanced. Hildreth and Houck (1968), for example, proposed estimators for the specification

\[ y_t = x_t^\prime \beta_t^* \]

(6)

where the first element of \( x_t \) is unity and the \( k \)-vector of coefficients \( \beta_t^* \) is distributed independently as \( N(\beta, \Lambda) \) for \( t = 1, \ldots, n \). Writing \( \beta_t^* = \beta + \nu_t \) implies that the \( \nu_t \) are \( \text{NID}(0, \Lambda) \), so that substituting in (6) would give

\[ y_t = x_t^\prime \beta + x_t^\prime \nu_t \]

which is the fixed coefficients model as in (1) with \( u_t = x_t^\prime \nu_t \) as the disturbance. Then \( u_t \sim \text{NID}(0, \sigma_t^2) \) with

---

3 Hildreth and Houck (1968) restricted their analysis to a diagonal \( \Lambda \) matrix but later generalizations including Swamy (1970) have a full \( \Lambda \) matrix as well as other features such as the same \( \beta_t^* \) realization as the coefficient vector for subgroups of observations.
\[ \sigma_t^2 = x_t' \Delta x_t \]

\[ = d_{11} + \sum_{i=2}^{k} d_{ii} x_{it}^2 + 2 \sum_{i=1}^{k} \sum_{j=1}^{i-1} d_{ij} x_{it} x_{jt} \]

where \( d_{ij} \) is the \((i,j)\)th element of \( \Delta \). This expression for the disturbance variance has the same structure as (5) above with \( m = 1 \) and with the elements of the vector \( x_t' x_t \) being formed from the distinct elements of \( x_t' x_t \).

Some mention should be made of heteroscedastic patterns which do not fit the general specification that is being considered here. One that may be found in many textbooks is \( \sigma_t^2 = \sigma^2 z_t \) for scalar \( z_t \) which is usually taken to be one of the regressors or its square. But this assumption does not allow the hypothesis of homoscedasticity to be tested by standard methods because there is no parametric restriction which gives the null hypothesis as a special case. It could, however, be made testable by generalizing to \( \sigma_t^2 = \sigma^2 z_t^Y \) which then fits (4) above. Another assumption which is not included is \( \sigma_t^2 = [E(y_t)]^2 = (x_t' \beta)^2 \) that was proposed by Theil (1951). When cast into the form of (2), this specification implies a functional relationship between the \( \alpha \) parameters and the \( \beta \) regression coefficients in such a way that there is no regression without heteroscedasticity.

### 6.3 Difficulties with Maximum Likelihood Estimation of Certain Heteroscedastic Schemes

Several references were given in the preceding section in which the disturbance variance specification \( \sigma_t^2 = (z_t' \alpha)^m \) has been advocated. In some applications, this heteroscedastic formulation is obtained as an implication of randomly varying regression coefficients, while in others it is proposed merely as a simple parametric model relating the disturbance variance to a set of exogenous variables. However this model presents a
number of theoretical and practical difficulties for estimation and hypothesis testing, some of which appear to have been unrecognized by proponents of the specification.

One difficulty arises in the random coefficient model where $m = 1$ and the elements of the $\alpha$ vector are subject to certain prior restrictions related to their interpretation as variances and covariances. This problem has been given extensive consideration, for example Swamy (1971, pp.107-110) discusses in a general framework the meaning and treatment of negative variance estimates. Dent and Hildreth (1977) compare several algorithms for maximum likelihood estimation of the random coefficient regression model with diagonal $\Delta$ matrix, including enforcement of non-negativity constraints on the estimates of the diagonal elements. In this situation, testing the hypothesis of no heteroscedasticity (fixed coefficients) involves parameter values on the boundary of the parameter space, so that LR and W statistics would not have the usual $\chi^2$ asymptotic distributions; but as noted in §1.7, the LM test will have the usual properties.

Another variation on the standard situation arises when $m$ is an even integer (usually $m = 2$). With this assumption the signs of the $\alpha$ coefficients will not be identifiable because, for any value of $\alpha$, both $+\alpha$ and $-\alpha$ give exactly the same $\sigma^2_t$ for each observation and therefore correspond to the same likelihood function value. Both Rutemiller and Bowers (1968) and Goldfeld and Quandt (1972, Ch. 3) propose numerical algorithms for full maximum likelihood estimation of the regression model with disturbance variance $\sigma^2_t = (z'_t \alpha)^2$, but neither study mentions the identification problem. In any case, the implications do not appear to be serious, as local identifiability is possible under certain conditions and inference is not affected by the sign of $\alpha$ since both $+\hat{\alpha}$ and $-\hat{\alpha}$ would give the same $W$ and LR statistics for testing homoscedasticity.
Another difficulty with this specification is that the likelihood function may be unbounded. This problem arises because of singularities in the likelihood function associated with zero variance estimates and, while values of \( \alpha \) which give \( z_t'\alpha = 0 \) for any \( t \) can and should be excluded from the admissible parameter space, neighbouring \( \alpha \) values can give rise to arbitrarily large likelihood values. To see this, consider the regression model (1) with disturbance variance \( \sigma_t^2 = (z_t'\alpha)^m \) where \( m \) is a prespecified integer. Suppose the parameter space of \( \alpha \) is defined by those values for which \( z_t'\alpha > 0 \) for all \( t = 1, \ldots, n \). When \( m \) is odd this condition is the same as \( \sigma_t^2 > 0 \) and when \( m \) is even it also serves to include an identifiability condition on the sign of the \( \alpha \) vector. The likelihood can be written as

\[
(2\pi)^{-\frac{n}{2}} \sigma_j^{-1} \exp\{-\frac{1}{2} \sigma_j^{-2} (y_j - x_j')^2\} \prod_{t=1}^{n} \sigma_t^{-1} \exp\{-\frac{1}{2} \sigma_t^{-2} (y_t - x_t')^2\}.
\]

Then if there exists a value of \( \alpha \) (say \( \alpha_o \)) on the edge of, but not included in, the parameter space such that \( z_j'\alpha_o = 0 \) while \( z_t'\alpha_o > 0 \) for all \( t \neq j \), it will be possible to consider a sequence of values for \( \alpha \) within the parameter space such that \( \sigma_j^2 \to 0 \) as \( \alpha \to \alpha_o \) while all other \( \sigma_t^2 \) are bounded away from zero. With \( \beta \) chosen so that \( y_j - x_j'\beta = 0 \), the likelihood will increase without bound as \( \alpha \to \alpha_o \).

Unboundedness of the likelihood function does not imply a failure of the method of maximum likelihood in the sense that, for sufficiently large \( n \), there will exist a consistent root of the likelihood normal equations giving a well-behaved asymptotically normal estimator, in problems that are otherwise regular [Kiefer and Wolfowitz (1956)]. In practice, however, difficulties may be experienced when unsuitable initial estimates in the neighbourhood of a singularity point are used in a maximization algorithm.
For some of the heteroscedastic alternative hypotheses that have been proposed in the literature, maximum likelihood estimation of the full model may present considerable problems. The LM test which is obtained in the next section does not require estimation of the alternative hypothesis generalization and thus avoids these difficulties.

6.4 Derivation of the LM Statistic for Heteroscedasticity

For some purposes, it will be more convenient to write all \( n \) observations on the regression model as

\[
y = X\beta + u
\]

in the usual notation, and to define \( Z \) as the \( n \times (p+1) \) matrix which has \( z_t' \) as its \( t' \)th row. Maximum likelihood estimates under the null hypothesis that the disturbance is homoscedastic will be given by applying OLS to (7); these are denoted by \( \hat{\beta} \) and \( \hat{\sigma}^2 \), giving residuals

\[
\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_n)' = y - X\hat{\beta}
\]

so that \( \hat{\sigma}^2 = n^{-1}\tilde{u}'\tilde{u} \). Also useful will be the quantities

\[
\tilde{g}_t = \tilde{u}_t^2/\hat{\sigma}^2 \quad \text{and} \quad \tilde{\ell}_t = (\tilde{u}_t^2/\hat{\sigma}^2 - 1) = (\tilde{g}_t - 1)
\]

and the corresponding vectors

\[
\tilde{g} = (\tilde{g}_1, \ldots, \tilde{g}_n)'
\]

\[
\tilde{\ell} = (\tilde{\ell}_1, \ldots, \tilde{\ell}_n)' = \tilde{g} - \ell
\]

where \( \ell \) is an \( n \)-vector of units. Note that

\[
\ell'\tilde{g} = n \quad \text{and} \quad \ell'\tilde{\ell} = 0
\]

and that \( \ell \) is the first column of \( Z \) because each \( z_t \) has a unit as its first element.
Under the assumptions of §6.2, the log-likelihood for the unknown parameters in the alternative hypothesis heteroscedastic generalization is

$$l(\alpha, \beta) = -\frac{1}{2} n \log (2\pi) - \frac{1}{2} \sum_t \log \sigma_t^2 - \frac{1}{2} \sum_t u_t^2 / \sigma_t^2$$

where $$u_t = y_t - x_t^T \beta$$ and $$\sigma_t^2 = h(z_t^T \alpha)$$. Now it is not difficult to show that the information matrix is block-diagonal between $$\alpha$$ and $$\beta$$ parameters, so that only the components of the score vector and information matrix relating to $$\alpha$$ will be required. These are

$$d_{\alpha}(\alpha, \beta) = \frac{1}{2} \sum_t \left[ \frac{h'(z_t^T \alpha)}{\sigma_t^2} \right] z_t \left( \frac{u_t^2}{\sigma_t^2} - 1 \right)$$

where $$h'(s) = \partial h(s)/\partial s$$, and using

$$\frac{\partial^2}{\partial \alpha \partial \alpha^T} = -\frac{1}{2} \sum_t \left\{ \left[ \frac{h'(z_t^T \alpha)}{\sigma_t^2} \right]^2 z_t z_t^T \left( 2\frac{u_t^2}{\sigma_t^2} - 1 \right) - \left[ \frac{h''(z_t^T \alpha)}{\sigma_t^2} \right] z_t z_t^T \left( \frac{u_t^2}{\sigma_t^2} - 1 \right) \right\}$$

where $$h''(s) = \partial^2 h(s)/\partial s^2$$, gives

$$I_{\alpha \alpha} = \frac{1}{2} \sum_t \left[ \frac{h'(z_t^T \alpha)}{\sigma_t^2} \right]^2 z_t z_t^T$$

since $$E(u_t^2/\sigma_t^2) = 1$$.

Evaluating the score vector and information matrix at the restricted (OLS) estimates then gives

$$d_{\alpha}(\tilde{\alpha}, \tilde{\beta}) = \frac{1}{2} \sum_t \left[ \frac{h'(\tilde{\alpha}_1)}{\tilde{\sigma}_t^2} \right] z_t \tilde{z}_t$$

$$\tilde{I}_{\alpha \alpha} = \frac{1}{2} \sum_t \left[ \frac{h'(\tilde{\alpha}_1)}{\tilde{\sigma}_t^2} \right]^2 z_t \tilde{z}_t$$
so that the LM criterion for testing the hypothesis is obtained as

\[
LM = \left[ d_{\alpha} (\hat{\alpha}, \hat{\beta}) \right]' T^{-1} \left[ d_{\alpha} (\hat{\alpha}, \hat{\beta}) \right]
\]

\[
= \frac{1}{2} \left( \sum_t z_t \hat{r}_t \right)' \left( \sum_t z_t z_t' \right)^{-1} \left( \sum_t z_t \hat{r}_t \right)
\]

In forming the LM statistic, note that the factor \([h'(\hat{\alpha})/\sigma^2]\) appearing in both the score vector and information matrix will cancel, so the statistic will not depend upon the particular functional form of \(h(\cdot)\).

Using the vector notation, the criterion becomes

\[
LM = \frac{1}{2} \hat{r}' Z(Z'Z)^{-1} Z' \hat{r}
\]

\[
= \frac{1}{2} [\hat{g}' Z(Z'Z)^{-1} Z' \hat{g} - n]
\]

\[
= \frac{1}{2} [\hat{g}' Z(Z'Z)^{-1} Z' \hat{g} - n^{-1}(\hat{\iota}' \hat{\iota})^2]
\]

where the second line follows from substituting \(\hat{r} = (\hat{g} - \hat{\iota})\) and noting that \(Z(Z'Z)^{-1} Z' \hat{\iota} = \hat{\iota}\) because \(\hat{\iota}\) is the first column of \(Z\). Expression (8) would be one half of the explained sum of squares from the regression of \(\hat{r}_t = (\hat{u}_t/\sigma^2 - 1)\) upon \(z_t\) while (9) would be a similar quantity from the regression of \(\hat{g}_t = \hat{u}_t/\sigma^2\) upon \(z_t\), and either of these may be used as a simple method of computing the statistic. From the general asymptotic theory, the LM test would be performed by rejecting the null hypothesis of homoscedasticity when (8), or equivalently (9), exceeds the appropriate upper point of the \(\chi^2(p)\) distribution where there are \(p\) constraints imposed by the hypothesis.

In some cases it may be possible to simplify the statistic even further. For example, if it is postulated that the disturbance variance undergoes a discrete change after \(n_1\) of the \(n\) observations, then the appropriate \(Z_{2t}\) would be a dummy variable taking one value (say unity) for the first \(n_1\) observations and another value (say zero) for the latter
\( n_2 = (n-n_1) \) observations. Then

\[
\begin{bmatrix}
  n_1 & n_1 \\
  n_1 & n_1
\end{bmatrix}
\begin{bmatrix}
  Z'Z \\
  Z'\tilde{f}
\end{bmatrix} =
\begin{bmatrix}
  0 \\
  n_1 \\
  \Sigma \tilde{f}_t \\
  t = 1, \ldots, n
\end{bmatrix}
\]

and

\[
LM = k \frac{n}{n_2 n_1} \left( \frac{n_1}{\Sigma \tilde{f}_t} \right)^2
\]

\[
= k \frac{n_1 n_2}{n} \left( \frac{\bar{\sigma}_1^2 - \bar{\sigma}_2^2}{\bar{\sigma}_2^2} \right)^2
\]

where \( \bar{\sigma}_1^2 = n_1^{-1} \Sigma \bar{u}_t^2 \) and \( \bar{\sigma}_2^2 = n_2^{-1} \Sigma \bar{u}_t^2 \) are estimates of the variance in the first and second subsamples respectively, using residuals from one regression on all \( n \) observations.

It is interesting that the expression for the LM statistic does not depend upon the particular functional form \( h(\cdot) \) by which the linear combination \( z^t a \) determines the heteroscedastic variance under the alternative hypothesis. As was argued in Chapter 3, this invariance of the LM criterion to specific details of the alternative hypothesis generalization may be a desirable feature for tests of misspecification to have. In practice, a researcher may be able to suggest in general terms the nature and source of potential misspecification, e.g. heteroscedasticity related to some \( z_t \) variables, but rarely would prior knowledge be as specific as detailing the precise functional form of the relationship. Even if a specific alternative is suggested, the computational simplicity of the LM test relative to other asymptotically equivalent procedures may make its usage attractive.
6.5 Finite-Sample Properties

As with all procedures developed from asymptotic principles, it is of some interest to investigate the properties of the LM test for heteroscedasticity in finite samples. The most interesting questions would focus on the adequacy of the $\chi^2$ distribution as an indicator of significance probabilities and the power of the test to reject false null hypotheses compared with other procedures such as W or LR. To this end, a series of Monte Carlo simulation experiments were performed and these are reported in Chapter 7, but firstly it is desirable to pursue analytical methods as far as practicable. The analysis in this section considers only the situation in which the null hypothesis is correct, although the methods that are used could also be applied to alternative hypotheses.

For a discussion of finite-sample properties, some rearrangements of the statistic will be useful. Define

$$A = Z(Z'Z)^{-1}Z' \quad \text{and} \quad Q = I_n - X(X'X)^{-1}X'$$

giving the OLS residuals as $\tilde{u} = Qu$ and the criterion as

$$LM = \frac{1}{2} [\tilde{g}'A\tilde{g} - n] . \tag{10}$$

Alternatively, let $\tilde{v}$ be the vector of squared OLS residuals, i.e.

$$\tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_n)' \quad \text{with} \quad \tilde{v}_t = \tilde{u}_t^2 \quad \text{for} \quad t = 1, \ldots, n,$$

so that $\tilde{v} = \tilde{o}^2g$ and

$$LM = \frac{n}{2} \frac{\tilde{v}'(A-\tilde{\iota}(\tilde{\iota}'\tilde{\iota})^{-1}\tilde{\iota}')\tilde{v}}{\tilde{v}'(\tilde{\iota}'\tilde{\iota})^{-1}\tilde{\iota}'\tilde{v}} . \tag{11}$$

In this form, it can be seen that the LM statistic is a ratio of idempotent quadratic forms in the squared OLS residuals. These squared residuals are dependently distributed gamma random variables and very little attention appears to have been given to the distribution of
quadratic forms or ratios of quadratic forms in such quantities.

(i) \textbf{Moments}

One possible approach to obtaining some information about the exact finite-sample distribution of the statistic is via the moments for which expressions can, in principle, be derived. The first few moments, while yielding only partial information about the distribution itself, could be used as a check on the asymptotic distribution as an approximation to the actual distribution in finite samples. If the limiting $\chi^2$ approximation is found to be inadequate, moments could be used to obtain better approximations, say by fitting a simple empirical distribution such as the beta or some other member of the Pearson system, or more generally in an Edgeworth expansion. Thus, if moments can be calculated fairly simply, they may give useful information for checking and improving upon asymptotic results.

From (10),

$$LM = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \tilde{g}_i \tilde{g}_j - n$$

where $A = \{a_{ij}\}$ for the individual elements. Now, using $e_i$ to represent the $i$'th column of the $n \times n$ identity matrix,

$$\tilde{g}_i \tilde{g}_j = [\tilde{u}_i \tilde{u}_j / \sigma^2]^2 = \frac{1}{2} n^2 [u'Q(e_i e_j' + e_j e_i')Qu/u'Qu]^2$$

$$= \frac{1}{2} n^2 [N/D]^2$$

where $\tilde{u}_i = e_i' \tilde{u} = e_i' Qu$ and where $N$ and $D$ are used as shorthand for numerator and denominator quadratic forms respectively. With $Q$ being idempotent, $(e_i e_j' + e_j e_i')$ being symmetric and $u_t \sim NID(0, \sigma^2)$ under the null hypothesis, the ratio $N/D$ will be distributed independently of the denominator $D$, by a result attributed to E.J.G. Pitman. [See Hannan (1970,
This result implies, in particular,

\[ E(N)^S = E[(N/D)^S(D)^S] = E(N/D)^S E(D)^S \]

so that

\[ E(N/D)^S = E(N)^S / E(D)^S \]

for any \( s \), giving for \( \tilde{g}_{ij} = \tilde{u}_{ij} / \tilde{\sigma}^2 \)

\[ E(\tilde{g}_{ij} \tilde{g}_{ij})^S = E(\tilde{u}_{ij} \tilde{u}_{ij})^S / E(\tilde{\sigma}^2)^{2S} \]

Moments of numerator and denominator separately, when \( H_0 \) is correct, can then be obtained from

\[ \tilde{u} \sim N(0, \sigma^2 Q) \quad \text{and} \quad n\tilde{\sigma}^2 / \sigma^2 \sim \chi^2(n-k) \]

For the mean of the LM statistic,

\[ E(\tilde{u}_{ij} \tilde{u}_{ij}) = \sigma^4 (q_{ij}^2 + 2q_{ij}^2) \]

\[ E(\tilde{\sigma}^2)^2 = \sigma^4 n^{-2} [2(n-k) + (n-k)^2] \]

where \( Q = \{ q_{ij} \} \), giving

\[ E(LM) = \frac{1}{2} \left[ \sum_i \sum_j a_{ij} E(\tilde{g}_{ij} \tilde{g}_{ij}) - n \right] \]

\[ = \frac{1}{2} \left[ \frac{n^2}{2(n-k)+(n-k)^2} \sum_i \sum_j a_{ij} (q_{ij}^2 + 2q_{ij}^2) - n \right]. \]

Calculating the second moment of the LM statistic in a similar manner would involve:

\[ E(\tilde{u}_{ij} \tilde{u}_{kl} \tilde{u}_{ij} \tilde{u}_{kl}) = \sigma^8 \left[ q_{ii}^2 q_{jj}^2 q_{kk}^2 q_{ll}^2 q_{kkll}^2 \right. \]

\[ + 2(q_{ij}^2 q_{kl}^2 + q_{ii}^2 q_{kk}^2 q_{ll}^2 + q_{ii}^2 q_{kkll}^2 + q_{ij}^2 q_{kkll}^2 + q_{jj}^2 q_{kkll}^2 + q_{ij}^2 q_{kkll}^2 + q_{kl}^2 q_{kkll}^2 + q_{kkll}^2 q_{ij}^2) \]
+ 4(q_{ij}^2q_{kl}^2 + q_{ik}^2q_{lj}^2 + q_{il}^2q_{jk}^2)
+ 8(q_{ij}q_{jk}q_{kl}q_{lj} + q_{ij}q_{ik}q_{lk}q_{lj} + q_{ij}q_{il}q_{jk}q_{lk} + q_{ij}q_{ik}q_{il}q_{jk})
+ 16(q_{ij}q_{jk}q_{kl}q_{li} + q_{ij}q_{ik}q_{lk}q_{li} + q_{ij}q_{il}q_{jk}q_{li} + q_{ij}q_{ik}q_{il}q_{jk})

E(\hat{\sigma}^2)^4 = \sigma^4 n^{-4} [48(n-k) + 44(n-k)^2 + 12(n-k)^3 + (n-k)^4]

which would give a very much more complicated expression for \( E(LM)^2 \) or \( \text{var}(LM) \) than for the mean.

Apart from showing that the distribution of the statistic will depend in a complicated way upon all of the \( x_t \) and \( z_t \) exogenous observations, proceeding via the moments does not appear to be very practicable in this instance. A number of avenues were explored of approximating the exact moments by including second-order terms (i.e. those which are \( O(n^{-1}) \)), or by using the properties of matrices \( \Lambda \) and \( Q \) to obtain bounds on the mean independently of the particular set of \( x_t \)'s and \( z_t \)'s. However, none of these attempts produced expressions simple enough for practical usage.

(ii) **Numerical Methods**

Some simplifications are possible when \( p = 1 \), i.e. when \( z_t'\alpha = \alpha_1 + \alpha_2 z_t \) with scalar \( Z_t \). This is an important special case because most practical applications of the test for heteroscedasticity, except when it derives from a random coefficient hypothesis, would have the disturbance variance related to only one exogenous variable. When \( p = 1 \), the rank of \( \Lambda \) is 2 and the rank of the matrix in the numerator quadratic form of (11) will be \( \text{rank}(\Lambda) - 1 = 1 \), so that this matrix can be written as

\[ \Lambda - \dot{\iota}(\dot{\iota}'\dot{\iota})^{-1}\dot{\iota}' = bb' \]
where \( b \) is an \( n \)-vector with, as its \( i \)'th element,

\[
b_i = \frac{(Z_i - \bar{Z})}{\left[ \sum_{t=1}^{n} (Z_t - \bar{Z})^2 \right]^{\frac{1}{2}}}
\]

where \( \bar{Z} = \frac{1}{n} \sum_{t} Z_t \). In this case, expression (11) becomes

\[
LM = \frac{n}{2} \frac{\bar{v}' \bar{b} \bar{v}}{\bar{v}' \bar{v}}
\]

\[
= \frac{n}{2} \left( \frac{\bar{b}' \bar{v}}{\bar{v}' \bar{v}} \right)^2
\]

\[
= \left( \frac{\bar{u}' \bar{D} \bar{u}}{\bar{u}' \bar{u}} \right)^2
\]

(12)

where \( D \) is a diagonal \( n \times n \) matrix with \( \frac{nb_i}{\sqrt{2}} \) as its \( i \)'th diagonal element, for \( i = 1, \ldots, n \).

One application of (12) is to enable exact significance probabilities to be computed by standard techniques, for a given set of exogenous observations. For any \( c > 0 \),

\[
Pr(LM > c) = Pr(\bar{u}' \bar{D} \bar{u} / \bar{u}' \bar{u} > \sqrt{c}) + Pr(\bar{u}' \bar{D} \bar{u} / \bar{u}' \bar{u} < -\sqrt{c})
\]

(13)

\[
= P_1 + P_2.
\]

Now,

\[
\bar{u}' \bar{D} \bar{u} / \bar{u}' \bar{u} = u'Q Du / u'Qu
\]

is a ratio of quadratic forms in normally distributed random variables which, under the null hypothesis, can be taken to be \( u_t / \sigma \) which are NID(0,1). Given \( Q \) which is a function of the observed \( x_t \) exogenous variables and \( D \) which depends on the \( z_t \) observations, probabilities \( P_1 \) and \( P_2 \) can be computed using numerical inversion of the characteristic function by numerical integration, as detailed in Imhof (1961).
Unlike the somewhat similar application to the Durbin-Watson statistic, for which Koerts and Abrahamse (1969) employed the Imhof method, there does not appear to be any basis for attaching a sign to $\sqrt{LM}$ to give a test of a one-sided alternative hypothesis. Therefore in this application, two probabilities would have to be computed and, with two numerical integrations to be performed, the exact procedure will tend to be computationally expensive. An alternative approach, which extends to $p > 1$ and which, for most purposes, is comparable in efficiency and economy with the Imhof method, is to estimate the exact probabilities by simulation. Observe that, under the null hypothesis, the finite-sample distribution of the LM statistic will not depend upon any unknown parameters so only one experiment is required for a given set of exogenous observations. If $P = \Pr(LM > c)$ is required for some $c$, this may be estimated by:

(i) generating $m$ sets of $n$ observations on a pseudorandom NID(0,1) variable as $u_t$, (ii) forming the statistic as in any of the formulae given above with $\hat{u} = Qu$, (iii) counting the number of times, $r$, that $LM > c$ in these $m$ sets, and (iv) using $r/m$ as an estimate of $P$.

Comparisons of relative computational requirements between simulation and the numerical integration method (for $p = 1$) are difficult because computer time with the latter method was found to vary markedly (up to a factor of 3) with the particular set of $z_t$'s and the extremity into the tail of the distribution being examined. On average, $m = 5000$ in simulation was found to be roughly comparable with exact computations, and this would allow $P = .05$ to be estimated with standard error of $[P(1-P)/5000]^{1/2} = .003$, which should be more than adequate accuracy for most applications. Even when $p = 1$ and Imhof's method is available, it should be borne in mind that errors in numerical integration may be as high as those from simulation if the truncation point is too small or the grid is too large.
(iii) **Bounds Tests**

Even though exact significance probabilities can be obtained or estimated closely, it seems likely that most researchers would use this approach only if the computed value of the test statistic was near the critical point taken from the asymptotic $\chi^2$ distribution, or if the limiting approximation was found to be a poor guide to the exact distribution in finite samples. In the case of the Durbin-Watson statistic, the method of Koërts and Abrahamse is little used in practice, but an exact solution would be required here only if the statistic happened to lie between the bounds set by extremes of regressor configurations. Analogous bounds for the heteroscedasticity test would depend upon the set of $z_t$ observations, so tabulation of bounds for this criterion would not be a useful procedure. (In the Durbin-Watson statistic, the matrix corresponding to $D$ in (12) is constant and independent of the data.) However, for a given $Z_t$ set in the case $p = 1$, i.e. $z_t' \alpha = \alpha_1 + \alpha_2 Z_t$, it would be possible to obtain bounds which encompass the effects of all possible sets of regressors. Computation of bounds on probabilities may be a useful approach to analyzing some problems, such as the following one.

From (13), observe that for critical values, $c$, greater than

$$c^* = \frac{1}{2} n^2 \max_i \left\{ \left( Z_i - \bar{Z} \right)^2 / \sum_{t=1}^{n} \left( Z_i - \bar{Z} \right)^2 \right\}$$

both $(D - \sqrt{c} I_n)$ and $(-D - \sqrt{c} I_n)$ will be negative definite so that

$$P = \Pr(\text{LM} > c) = P_1 + P_2$$

will be identically zero. Since the approximating $\chi^2$ distribution has nonzero probabilities in the tail beyond any $c$, might not the $\chi^2$ give misleading indications of significance probabilities for certain sorts of $Z_t$'s, particularly as it is the tail of the distribution which is of interest for hypothesis testing?
To examine this question, firstly note that

\[
\max_i (Z_i - \bar{Z})^2 \geq n^{-1} \sum_{t=1}^n (Z_t - \bar{Z})^2
\]

so that the point \( c^* \) at which the distribution of \( LM \) is truncated is at least as large as \( \frac{1}{2} n \). Typical values of \( c \) that would be of interest would generally be less than about 7, corresponding to significance levels greater than about 0.01 in the \( \chi^2(1) \), so it is only with very small samples that actual truncation would be a problem at the significance levels that are likely to be used. However, the truncated nature of the distribution of \( LM \) does suggest that the asymptotic \( \chi^2(1) \) may systematically understate the true significance of the \( LM \) criterion, particularly for low significance levels, smaller sample sizes and certain sorts of \( Z_t \)'s.

The worst kind of \( Z_t \) in this respect would be one for which the equality in (14) holds, implying that \( (Z_t - \bar{Z})^2 \) is constant for all observations and giving \( c^* = \frac{1}{2} n \). This would be the situation when the disturbance variance is hypothesized to take one value for half of the observations and another value for the other half, for which the appropriate \( Z_t \) is a dummy variable. With some suggestion that such a \( Z_t \) might be associated with large discrepancies between asymptotic theory and small sample behaviour, placing bounds upon the effects of the regressors seems to be a useful way to analyze the situation.

The working here closely follows that of Durbin and Watson (1950) - with the corrections in Durbin and Watson (1951) - except that the criterion in the heteroscedasticity test is the square of a ratio of quadratic forms and not simply the ratio itself as it is in the autocorrelation test. For notational convenience, let \( d = \sqrt{LM} \) so that
\[ d = \frac{\tilde{u}'D\tilde{u}}{\tilde{u}'\tilde{u}} = \frac{u'QDQu}{u'Qu} \]

and let \( \lambda_1, \ldots, \lambda_n \) be the eigenvalues of \( D \), with indexing arranged so that

\[ \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \]

Since \( D \) is diagonal its eigenvalues are its diagonal elements, and with the dummy variable \( Z_t \),

\[
\lambda_i = \begin{cases} 
- \sqrt{\frac{i}{n}} & \text{for } i = 1, \ldots, \frac{n}{2} \\
\sqrt{\frac{i}{n}} & \text{for } i = (\frac{n}{2} + 1), \ldots, n 
\end{cases} \quad (15)
\]

Denote the ordered nonzero eigenvalues of \( QD \) (which are the same as those of \( QDQ \)) by

\[ \nu_1 \leq \nu_2 \leq \ldots \leq \nu_{n-k} \]

where \( k \) of them are zeros because \( \text{rank}(QD) = \text{rank}(Q) = (n-k) \).

Now the discussion in Durbin and Watson (1950, pp.415-416) is sufficiently general to give for the present case:

(i) there is an orthogonal transformation from \((u/\sigma)\) to \( w = (w_1, \ldots, w_n)' \), implying \( w \sim N(0, I_n) \), such that

\[
d = \frac{\sum_{i=1}^{n-k} \nu_i w_i^2}{\sum_{i=1}^{n-k} w_i^2} ; \quad (16)
\]

(ii) \( \lambda_1 \leq \nu_i \leq \lambda_i + k \) for \( i = 1, \ldots, (n-k) \); \quad (17)

\[ ^4 \text{Durbin and Watson (1950, 1951) also allow for the possibility of providing tighter bounds when the eigenvectors of } D \text{ (in this case the columns of the } n \times n \text{ identity matrix) are linear combinations of the regressors, but this refinement is not warranted for the present purpose.} \]
(iii) there exist bounding random variables giving \( d_L \leq d \leq d_u \),

where

\[
    d_L = \frac{\sum_{i=1}^{n-k} \lambda_i w_i^2}{\sum_{i=1}^{n-k} w_i},
\]

\[
    d_u = \frac{\sum_{i=1}^{n-k} \lambda_{i+k} w_i^2}{\sum_{i=1}^{n-k} w_i}.
\]

These bounds apply to both \( P_1 = \Pr(d > \sqrt{c}) \) and \( P_2 = \Pr(d < -\sqrt{c}) \) individually, but for \( P = P_1 + P_2 \) which is of interest here, the issue is not so straightforward. Clearly, if the extreme \( \lambda_i \) are chosen as in \( d_u \) so as to set \( P_1 \) at a maximum then \( P_2 \) will be at a minimum and vice versa. However, from (15) and (17),

\[
    v_1 = -\sqrt{\frac{1}{2}n} \quad \text{for} \quad i \in i_1 \equiv \{ i \mid i = 1, \ldots, (\frac{1}{2}n-k) \}
\]

\[
    -\sqrt{\frac{1}{2}n} \leq v_1 \leq \sqrt{\frac{1}{2}n} \quad \text{for} \quad i \in i_2 \equiv \{ i \mid i = (\frac{1}{2}n-k+1), \ldots, \frac{1}{2}n \}
\]

\[
    v_1 = \sqrt{\frac{1}{2}n} \quad \text{for} \quad i \in i_3 \equiv \{ i \mid i = (\frac{1}{2}n+1), \ldots, (n-k) \}
\]

so that the problem of finding upper and lower bounds on \( P \) amounts to determining \( v_1 \) for \( i \in i_2 \) in the range \(-\sqrt{\frac{1}{2}n} \leq v_1 \leq \sqrt{\frac{1}{2}n}\). From (16), using \( i_4 \equiv \{ i \mid i = 1, \ldots, (n-k) \} = i_1 \cup i_2 \cup i_3 \),

\[
    P = \Pr(LM > c) = \Pr \left( \frac{\sum_{i \in i_2} v_1 w_i^2 + \eta}{\sum_{i \in i_4} w_i^2} > c \right)
\]

where \( \eta = \sqrt{\frac{1}{2}n} \left( \sum_{i \in i_3} w_i^2 - \sum_{i \in i_1} w_i^2 \right) \) is symmetrically distributed about the origin, independently of \( \sum_{i \in i_2} v_1 w_i^2 \). Then it is clear that \( P \) is minimized by setting \( v_1 = 0 \) for all \( i \in i_2 \) and it is maximized equally by both \( v_1 = \sqrt{\frac{1}{2}n} \) for all \( i \in i_2 \) and \( v_1 = -\sqrt{\frac{1}{2}n} \) for all \( i \in i_2 \).

Having determined the configuration of eigenvalues of \( QD \) which give upper and lower bounds on \( \Pr(LM > c) \), the exact tail probability
corresponding to any critical value, c, can be found by the Imhof numerical integration method. The highest and lowest probability that would be associated with six nominal significance points from the $\chi^2(1)$ asymptotic distribution for $k = 2$ and sample sizes 20, 40 and 80 are presented in the following table. It is felt that this arrangement is more informative for the present purpose of checking asymptotic theory in small to moderate samples than presenting bounding critical values for a given true significance level as in tables of the Durbin-Watson statistic.

<table>
<thead>
<tr>
<th>Critical Value</th>
<th>Nominal</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.642</td>
<td>.20</td>
<td>.1979</td>
<td>.2099</td>
<td></td>
</tr>
<tr>
<td>2.706</td>
<td>.10</td>
<td>.0960</td>
<td>.1051</td>
<td></td>
</tr>
<tr>
<td>3.841</td>
<td>.05</td>
<td>.0457</td>
<td>.0517</td>
<td></td>
</tr>
<tr>
<td>5.024</td>
<td>.025</td>
<td>.0212</td>
<td>.0249</td>
<td></td>
</tr>
<tr>
<td>6.635</td>
<td>.01</td>
<td>.0074</td>
<td>.0091</td>
<td></td>
</tr>
<tr>
<td>10.828</td>
<td>.001</td>
<td>.0004</td>
<td>.0006</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Critical Value</th>
<th>Nominal</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.642</td>
<td>.20</td>
<td>.1948</td>
<td>.2207</td>
<td>80</td>
</tr>
<tr>
<td>2.706</td>
<td>.10</td>
<td>.0912</td>
<td>.1106</td>
<td></td>
</tr>
<tr>
<td>3.841</td>
<td>.05</td>
<td>.0408</td>
<td>.0533</td>
<td></td>
</tr>
<tr>
<td>5.024</td>
<td>.025</td>
<td>.0172</td>
<td>.0244</td>
<td></td>
</tr>
<tr>
<td>6.635</td>
<td>.01</td>
<td>.0049</td>
<td>.0079</td>
<td></td>
</tr>
<tr>
<td>10.828</td>
<td>.001</td>
<td>.0001</td>
<td>.0002</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Critical Value</th>
<th>Nominal</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.642</td>
<td>.20</td>
<td>.1854</td>
<td>.2454</td>
<td>20</td>
</tr>
<tr>
<td>2.706</td>
<td>.10</td>
<td>.0793</td>
<td>.1233</td>
<td></td>
</tr>
<tr>
<td>3.841</td>
<td>.05</td>
<td>.0298</td>
<td>.0560</td>
<td></td>
</tr>
<tr>
<td>5.024</td>
<td>.025</td>
<td>.0092</td>
<td>.0218</td>
<td></td>
</tr>
<tr>
<td>6.635</td>
<td>.01</td>
<td>.0011</td>
<td>.0041</td>
<td></td>
</tr>
<tr>
<td>10.828</td>
<td>.001</td>
<td>.0000</td>
<td>.0000</td>
<td></td>
</tr>
</tbody>
</table>

From this table, it can be seen that the bounds converge with increasing sample size and that, for 40 observations or more, the asymptotic distribution would be a good guide, provided the significance probability is at least .05. With 20 observations, critical values from the $\chi^2(1)$ may be quite accurate for significance levels of .05 and larger, depending upon the regressors. For low nominal significance levels, .025 or smaller, asymptotic theory would systematically understate the true significance of
a calculated value of the LM statistic, whatever regressor set was used and even in samples as large as 80 observations. As expected, nominal \( \chi^2(1) \) probabilities become worse approximations to the true ones further into the tail. But for the usual significance levels that are employed in practice and for moderate samples, say \( n \geq 40 \), the truncated nature of the distribution of the LM statistic would not seem to present a serious problem to usage of the asymptotic \( \chi^2 \) approximation.
CHAPTER 7
A SIMULATION STUDY COMPARING W, LR AND LM TESTS
FOR HETEROSCEDASTICITY

7.1 Introduction

Some information about the behaviour in finite samples of the LM test for heteroscedasticity was obtained in Chapter 6. But these investigations were mainly concerned with examining the suggestion that the asymptotic $\chi^2$ distribution may be a poor indicator in small samples of the true significance of an observed value of the LM criterion in certain situations. Probably a more interesting question is: how does the LM test compare with other test procedures in situations similar to those in which it might be used in practice? The LM test shares certain asymptotic optimality properties with the W and LR tests, therefore it is of some considerable interest to examine the reliability of asymptotic results in the small samples that are typically available in practical applications.

There have been a number of advances made in the development of analytical methods for obtaining exact or approximate expressions for moments or distributions of econometric estimators, e.g. Nagar (1959), Sargan and Mikhail (1971), Phillips (1977). Application of similar methods to significance tests would generally involve higher order analysis than for the corresponding estimators [Sargan (1976, Sect. 4)], so it is not surprising that comparable analysis of tests has received relatively little attention. An advantage of the analytical approach is that it yields formulae which, while frequently very complicated expressions, do enable a wide range of situations to be predicted by substitution of parameter values into the formulae. While it may be desirable to pursue analytical methods for investigating the properties of statistical procedures as far as these
may be practicable, computer simulation experiments are often used to obtain some insights into the unknown properties in finite samples. One difficulty with simulation results is that they tend to be specific to the models which are studied, in contrast to the generality of analytic formulae. Also, by its very nature the simulation approach introduces stochastic experimental error into the outcomes, so the results of an experiment can at best be considered as estimates of the attributes of interest. Attention should therefore be given to the design and control of simulation experiments to reduce specificity and imprecision, and conclusions should be drawn accordingly.\footnote{An interesting discussion of the trade-off between the analytical and experimental approaches, as well as methods for combining information from both, is contained in Hendry (1975).}

A Monte Carlo study, designed to compare the LM test for heteroscedasticity with the corresponding W and LR tests, is reported in this chapter. In §7.2, the model that was studied is described together with the method that was used to obtain maximum likelihood estimates of the parameters. Design of the experiments and Monte Carlo methodology are discussed in §7.3 and §7.4 respectively, and in §7.5 the results are given and conclusions are drawn.

7.2 The Model and Estimation Procedure

The model investigated was the linear regression,

$$y_t = \beta_1 + \beta_2 x_{2t} + u_t$$

(1)

with multiplicatively heteroscedastic disturbance,

$$u_t \sim \text{NID}(0, \sigma_t^2) \text{ with } \sigma_t^2 = \sigma^2 \exp(\alpha_1 + \alpha_2 z_t)$$

(2)
where $\alpha_1 = \log \sigma_2^2$, $\alpha_2 = \gamma$ and $Z_t = \log(X_{jt})$. This model was chosen rather that the other common specification, $\sigma_t^2 = (z_t'\alpha)^m$, because of the problems that were discussed in §6.3, particularly that of an unbounded likelihood function. Some preliminary investigations with the models $\sigma_t^2 = (\alpha_1 + \alpha_2 Z_t)$ and $\sigma_t^2 = (\alpha_1 + \alpha_2 Z_t)^2$ using the artificial exogenous variable data described below found that the maximization algorithm would locate a singularity point quite frequently, even in samples with 80 observations and using the true parameter values as initial estimates.

With some exogenous variable data sets, "convergence" to a singularity was experienced in as many as 3% of the replications in preliminary simulations with these models. The situation is likely to be much worse when fewer observations are available, so a specification in which full estimation is relatively straightforward was selected for comparison of the LM test with W and LR procedures.

Maximum likelihood estimates of the parameters $(\beta_1, \beta_2, \alpha_1, \alpha_2)$ are required to perform W and LR tests. It will be convenient for discussion of the estimation method and the forms of the W and LR statistics to revert to the general notation of the preceding chapter with $x_t = (1, X_t)'$, $\beta = (\beta_1, \beta_2)'$, $z_t = (1, Z_t)'$ and $\alpha = (\alpha_1, \alpha_2)'$. Specializing the general derivations of §6.4 for the present heteroscedastic specification, the log-likelihood is

$$
\ell(\alpha, \beta) = -1/2 n \log(2\pi) - \frac{1}{2} \sum_t z_t' \alpha - \frac{1}{2} \sum_t \{exp(-z_t' \alpha)(y_t-x_t' \beta)^2 \} 
$$

with the first-order conditions for a maximum being

$$
\frac{\partial \ell}{\partial \beta} = 0 = \sum_t \{\text{exp}(-z_t' \alpha)\}x_t(y_t-x_t' \beta) 
$$

$$
\frac{\partial \ell}{\partial \alpha} = 0 = \frac{1}{2} \sum_t z_t\{\text{exp}(-z_t' \alpha)\}(y_t-x_t' \beta)^2 - 1 
$$
and the information matrix is

\[
I = \begin{bmatrix}
\sum_t \{\exp(-z_t^\prime \alpha)\} x_t x_t^\prime & 0 \\
0 & \frac{1}{2} \sum_t z_t z_t^\prime
\end{bmatrix}.
\]

Denote by \( \hat{\alpha} \) and \( \hat{\beta} \) the joint solution of (4) and (5) which maximizes the log-likelihood (3). Then from the first row of (5),

\[
\sum_t \{\exp(-z_t^\prime \alpha)\} (y_t - x_t^\prime \hat{\beta})^2 = n
\]

and substitution of this into (3) gives the supremum of the unrestricted likelihood as

\[
\max(\alpha, \beta) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \sum_t z_t^\prime \hat{\alpha} - \frac{1}{2} n.
\]

Under the null hypothesis of homoscedasticity, \( \alpha_2 = 0 \) and maximum likelihood estimates are given by OLS as

\[
\hat{\beta} = \left( \sum_t x_t x_t^\prime \right)^{-1} \sum_t x_t y_t
\]

\[
\hat{\sigma}^2 = n^{-1} \sum_t (y_t - x_t^\prime \hat{\beta})^2
\]

where \( \tilde{\alpha}_1 = \log \tilde{\sigma}^2 \) and \( \tilde{\alpha}_2 = 0 \). Substituting these into (3) gives the supremum of the likelihood, when it is constrained by the null hypothesis, to be

\[
\max(\tilde{\alpha}, \tilde{\beta}) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} n \log \tilde{\sigma}^2 - \frac{1}{2} n.
\]

From the definitions in §1.3, the LR statistic is then

\[
LR = 2[\max(\tilde{\alpha}, \tilde{\beta}) - \max(\hat{\alpha}, \hat{\beta})] = n \log \tilde{\sigma}^2 - \sum_t z_t \hat{\alpha} \tag{6}
\]
and the $W$ statistic is

$$W = \frac{1}{2} \frac{\hat{\alpha}_2^2}{v(Z)}$$  \hspace{1cm} (7)$$

where $v(Z)$ is the $(2,2)$ element of $\left( \sum_t z_t z_t' \right)^{-1}$.

The normal equations (4) and (5) are nonlinear in the parameters, thus making it impossible to obtain an explicit analytical solution for the unrestricted estimates. However equation (4) is linear in $\beta$ for a given value of $\alpha$ (rather, for a given $\alpha_2$, since $\alpha_1$ can be cancelled throughout). This suggests that standard nonlinear estimation algorithms might be simplified somewhat for the present application.

Analytical derivatives of the objective function (3) are readily obtained, so methods which use this information can be considered. Newton-Raphson, which is quadratically convergent, might be expected to perform well in this application; for functions such as the log-likelihood (3) which are not highly nonlinear, it can be expected to locate a maximum efficiently provided starting values are close to the maximum. [Bard (1974, Ch. V)]. The method of scoring has the advantage over Newton-Raphson for statistical applications that expected values are generally easier to compute than the second derivatives themselves and so this algorithm requires less computer time per iteration. In the present application, a large part of the information matrix is constant for all values of the parameters. Harvey (1976) suggested scoring as a suitable method for iterative estimation of multiplicative heteroscedasticity regression models, although his interest was mainly with two-step methods. Scoring has also been considered favourably for estimation of models which are somewhat similar: e.g., Swamy (1973) for random coefficient regressions and Harville (1977) for variance component models.

The structure of the problem suggests the following adaptation of the
scoring algorithm:

(i) with a trial value of $\alpha$, say $\alpha^S$, estimate $\beta$ as

$$\hat{\beta}^S = \left[ \sum_t \{ \exp(-z_t'\alpha^S) \} x_t x_t' \right]^{-1} \sum_t \{ \exp(-z_t'\alpha^S) \} x_t y_t$$

which will be given by the coefficients in a weighted regression of

$$\{ \exp(-\frac{1}{2}z_t'\alpha^S) \} y_t \text{ upon } \{ \exp(-\frac{1}{2}z_t'\alpha^S) \} x_t$$

(ii) form the residuals $\hat{u}_t^S = y_t - x_t'\hat{\beta}^S$ from (i), then compute

$$\hat{\alpha}^S = \left( \sum_t z_t z_t' \right)^{-1} \sum_t \{ \exp(-z_t'\alpha^S) \}(\hat{u}_t^S)$$

and update the estimate of $\alpha$ as $\alpha^{S+1} = \alpha^S + \hat{\alpha}^S - e_1$, where $e_1 = (1 \ 0)'$. Expression (9) is also a regression solution with $\{ \exp(-z_t'\alpha^S) \}(\hat{u}_t^S)^2$ being regressed upon $z_t$, and the dependent variable values for this regression can be obtained directly as the squares of the weighted residuals from step (i). Good convergence properties can be expected of this algorithm, which amounts to iterating the generalized least squares (Aitken) estimator. ² Although it is only possible in an experimental situation, the true values of $\alpha$ were used to initialize the iterations. Average numbers of iterations to convergence were approximately 5, 6 and 7 for 80, 40 and 20 observations respectively, although in some cases with 20 observations (to be noted in §7.5) the average was as high as 14.

Convergence criteria were set for both the change in the value of the likelihood function and the change in the estimates, from their values at

² Unlike many applications of iterated GLS, convergence is not guaranteed in this instance because step (ii) does not correspond to maximization over $\alpha$ for given $\hat{\beta}$. [See Oberhofer and Kmenta (1974).]
the previous iteration. The former was set in terms of the LR statistic and the latter by the W statistic such that

$$|\Delta LR| < .005 \quad \text{and} \quad |\Delta W| < .005$$

or

$$|\Delta LR/LR| < .001 \quad \text{and} \quad |\Delta W/W| < .001$$

defined convergence, where ΔLR and ΔW represent changes from the previous iteration. The relationship between convergence criteria in levels and in proportional changes was designed so that their stringency would be approximately equal in the region of the critical values of the test statistics, but when the accept vs reject decision was clear the weaker condition could prevail. Also, when W exceeded 10.0 so that the rejection decision was most unlikely to be changed by further iterations, the condition on W was relaxed to $$|\Delta W/W| < .01.$$³

Failures of the iterations to convergence by these (fairly weak) criteria were rare; generally less than 2 times in 1000 did failure occur with more than 40 observations. In some particular cases with 20 observations (noted in §7.5), the algorithm ran to 40 iterations without converging as many as 15 times in 1000. Progress of the algorithm was monitored for those situations where convergence did not occur by the above criteria and it was found that typically iterations were convergent but slow; divergent behaviour was observed but it was extremely rare, never occurring more than twice per 1000 replications.

³ Typically the convergence conditions on ΔLR were satisfied more readily than those on ΔW.
7.3 Experimental Design

It was noted in §6.5 that the LM statistic for testing heteroscedasticity has a null hypothesis distribution which does not depend upon the true values of the $\beta$ coefficients nor upon the scale of the disturbance variances. Distributions, both asymptotic and exact, will however depend upon the relative sizes of the variances at different observations under the alternative hypothesis. The exact distribution in finite samples, unlike the limiting asymptotic one, depends also upon the particular sets of $X$ and $Z$ exogenous observations that are used. It is noteworthy that the same conclusions apply also to the corresponding $W$ and LR statistics, although this is not so obvious because these criteria cannot be expressed in closed form as functions of the data. 

Invariance of the distributions of the three statistics to the values of certain parameters is useful information for designing Monte Carlo experiments which are intended to examine finite-sample properties of tests based upon these criteria. In particular, there is nothing to be gained by a design which incorporates more than one point in $\{\beta, \alpha_1\}$ space. The values $\beta_1 = \beta_2 = 0.0$ and $\alpha_1 = 1.0$ were most convenient for programming purposes and were used in the experiments, although the results are relevant to models with more "realistic" values of these parameters.

Dependence of the finite sample behaviour of the three tests upon the particular sets of $X$ and $Z$ exogenous observations presents considerable design difficulties. Even for the LM statistic which can be expressed in closed form as a function of disturbances and exogenous variables, it was noted in §6.5 that the distribution depends in a complicated manner on the values in $X$ and $Z$. It would be highly desirable

---

4 See Appendix C to this chapter.
to have some measure of the effects of the exogenous observations to use in design of this aspect of the experiments, but unfortunately no such measure suggests itself. As a practical but inferior alternative strategy, three different sets of artificial data were generated in a manner similar to previous investigations of heteroscedasticity. These data sets are denoted in the following sections as $X_{jt}$ ($j = 1, \ldots, 4$) where $X_{1t} = 1.0$ for all $t$, $X_{2t}$ was generated from a normal distribution with mean 50 and standard deviation 10, $X_{3t}$ was generated log-normally such that $\log(X_{3t}) \sim N(3,1)$ and $X_{4t}$ was generated from a uniform distribution on the interval $(1,31)$. Once generated, these exogenous variable observations were held constant for all experiments and are listed in Appendix B. Experiments with sample sizes 20, 40 and 80 used the first 20, the first 40 and all 80 observations on the exogenous variables. In all, 9 different combinations of regressor $X_{1t}$ in (1) and variance factor $X_{jt}$ in (2) were simulated for each sample size.

The one parameter in the model represented by (1) and (2) which does affect both asymptotic and finite-sample distributions of the test statistics is $\alpha_2$. To examine the adequacy of the asymptotic $\chi^2(1)$ distribution as an indicator of significance levels, the parameter value $\alpha_2 = 0$ representing the null hypothesis is of interest. Nonzero values of $\alpha_2$ are required to represent alternative hypotheses for examining questions of power.

Rather than attempting to estimate the whole of a power function, as done, e.g., in Mizon and Hendry (1979) for a test of dynamic specification,

---

E.g. Glejser (1969), Goldfeld and Quandt (1972, Ch. 3), Harvey and Phillips (1974). In §7.5 some attempts are made to explain variations between results obtained with different exogenous data sets using summary measures of the data.
it would appear in the present context to be more useful to examine just a few points on the power functions, while distinguishing between the various sets of exogenous observations. For each regressor, variance factor, sample size combination, two values of $\alpha_2$ were used with one representing an alternative hypothesis for which relatively low power might be expected and the other an alternative that would be easily detected. These values of $\alpha_2$ were determined in such a way that the asymptotic local power formula as in §1.5 would predict rejection probabilities of .20 and .80 respectively, given the nominal significance level which was set at .05.

When the null hypothesis is correct, each of the three statistics being examined here is asymptotically distributed as $\chi^2(1)$ so that the nominal .05 critical value is $c = 3.841$ according to

$$\Pr[\chi^2(1) > c] = .05.$$ 

To find the values of $\alpha_2$ for which the asymptotic theory would predict $P^* = .20$ and $P^* = .80$, it is necessary to find the value of the noncentrality parameter, $\phi$, such that

$$\Pr[\chi^2(1,\phi) > c] = P^*$$

where, from (7) above and the discussion of local power in §1.5,

$$\phi = \frac{1}{2} \frac{\alpha_2^2}{\nu(z)}.$$ 

Tabulations of $\phi$ for a range of $P^*$ values which includes $P^* = .80$ are given in Pearson and Hartley (1976) as Table 25, but unfortunately $P^* = .20$ is not included in the table. The noncentrality index which would give $P^* = .20$ in (10) was therefore obtained by trial-and-error solution with interpolation for $\phi$ in

$$\Pr[N(0,1) > (\sqrt{c} - \sqrt{\phi})] + \Pr[N(0,1) > (\sqrt{c} + \sqrt{\phi})] = .20$$
using the extended tables for the standard normal integral given in Pearson and Hartley (1976, Table 1). Having determined $\phi$ for each experiment according to whether it was the weak or the strong alternative hypothesis, the particular value given to $\alpha_2$ was then found according to the variance factor and the sample size, by

$$\alpha_2 = \sqrt{2} \phi \nu(Z)$$

where the positive square root was taken in every case. In the discussion of the results in §7.5, the actual values assigned to $\alpha_2$ are not reported. Since the main focus of interest is on how well asymptotic theory predicts the finite-sample behaviour of the three testing criteria in situations where discrimination between hypotheses is relatively easy and where it is relatively difficult, the actual values of $\alpha_2$ that were used are unimportant.

For descriptive purposes, an experiment is defined by the choice of the regressor variable (other than the constant), the variance factor (the $Z$ variable), the sample size and the level of heteroscedasticity. Thus there are $3 \times 3 \times 3 \times 3 = 81$ experiments.

7.4 Monte Carlo Technique

The experiments were all performed on the Australian National University UNIVAC 1100/42 computer with pseudorandom normal deviates generated by the subroutine GRAND. This routine is described, listed and evaluated in Brent (1974) and is a member of the class of algorithms reported favourably by Atkinson and Pearce (1976). A check was performed on 10000 sequential drawings from the generator by the Kolmogorov-Smirnov test which was unable to reject the hypothesis that they were observations on a $N(0,1)$ random variable, even at the .25 significance level. This
random number generating routine was judged to be adequate for the present application.

Experiments were conducted by generating a set of $N(0,1)$ realizations, $v_t$, for $t = 1, \ldots, n$, then computing

$$u_t = \exp\{1.0 + \alpha Z_t\} v_t$$

and forming

$$y_t = 0.0 + 0.0X_t + u_t.$$ 

Then, given $y_t$, $X_t$, $Z_t$ for $t = 1, \ldots, n$, the three statistics $W$, $LR$ and $LM$ were computed according to the estimation procedure and formulae in §7.2. A direct simulation approach in which this procedure was replicated 2000 times was used, with rejection probabilities estimated directly as the proportion of the 2000 replications in which the calculated values of the statistics exceeded the preset critical values.

The notation $\hat{P}$ is used for the estimate of a probability $P$, where $P^*$ is the prediction of $P$ from the asymptotic theory. In the experiments, $\hat{P}$ can differ from $P^*$ both because of discrepancies between $P$ and $P^*$ related to differences between finite-sample and asymptotic distributions and because of experimental error in estimating $P$ as $\hat{P}$.

By the direct simulation method, the standard error attached to an estimate of $P$ is given from the binomial distribution as $[P(1-P)/2000]^{1/2}$ which is approximately .005 for $P = .05$ and approximately .009 for $P = .20$ and $P = .80$. Thus, from the normal approximation to the binomial, individual discrepancies between an estimated probability and its asymptotic counterpart can be judged to be significant if they exceed .01 for null hypothesis situations and .02 for alternative hypothesis ones.

Standard errors on the differences between rejection probabilities
recorded for the three criteria can generally be expected to be considerably less than for individual estimates. Since all three statistics were computed using the same set of random numbers within an experiment and, as expected, the three statistics were found to be highly correlated across the 2000 replications, observed differences between the criteria are likely to be estimated better than individual rejection probabilities. Although it is difficult to formalize this notion in the context of estimating distributions, some justification is provided by analogy with the problem of estimating moments. The standard error attached to an estimate of the difference between the means of two random variables with the same variance will be less than the standard error on the estimate of an individual mean if the correlation between the two variables exceeds .5. For the different experiments, independent sets of random numbers were used. Comparing or averaging of results across similar but independent experiments can, therefore, be expected to highlight any systematic differences between the three testing criteria under study, while simultaneously reducing the effects of experimental stochastic errors.

The direct simulation approach is considered to be adequate for the present purpose although more sophisticated methods for simulating probability distributions are available. For example, Sargan (1976, Appendix D) describes a method whereby symmetric confidence intervals can be simulated using a control variable approach to increasing accuracy of estimates by reducing experimental error. Sargan's method for symmetric distributions could be adapted to the present situation, with the LM statistic for which the exact distribution is available being the obvious candidate for use as a control. However, the model being considered here is fairly simple and an efficient method is available for obtaining maximum likelihood estimates, so it is possible to use a large number of
replications at relatively low computational cost. It is doubtful that the increase in precision from a more sophisticated simulation method would justify the additional programming complexities required to implement it. If more accurate results were sought, it would probably be more efficient simply to increase the number of replications, especially as the numerical integrations for LM probabilities are computationally expensive anyway. In a sense, estimating LM probabilities together with the other two criteria rather than comparing exact LM probabilities with estimates of the others serves part of the purpose of a control variable. With the number of replications used in these experiments, simulation errors, especially on estimates of the differences between the three testing procedures, are controlled at a sufficiently low level for useful conclusions to be drawn in the next section.

In experiments with the model restricted by the null hypothesis, the 95'th percentile of the simulated values of each statistic was recorded as well as the proportion of values which exceeded the nominal .05 significance point. This was then used as an estimate of the true .05 significance critical value so that both the probability of a statistic exceeding the nominal critical value and power given a Type I error level of .05 could be estimated. Sampling error in the power estimates will of course be larger than in estimates of rejection probabilities for fixed critical values, so small observed differences should be interpreted accordingly.

7.5 Experimental Results

The results of the experiments are recorded in detail in Tables (i) to (ix) of Appendix A. Each table shows all of the outcomes for the nine experiments with one combination of regressor variable and disturbance
variance factor. Column headings denote whether a record corresponds to the null hypothesis, the weakly heteroscedastic alternative, $H_A^W$, or the strong one, $H_A^S$. Rejection probabilities for alternative hypotheses are further subdivided according to whether the nominal critical value was used (labelled "Reject") or the estimate of the true .05 significance point was used (labelled "Power"). Rows of the tables correspond to the three test criteria, grouped by sample size. Probabilities are recorded to three decimal places although the third place is not statistically significant for individual estimates, because more precise comparisons are possible within an experiment.

Ignoring for the moment the effects of different regressors and variance factors, averaging over all nine individual tables gives the following summary table.

<table>
<thead>
<tr>
<th>Nominal</th>
<th>$H_0$ Reject</th>
<th>$H_A^W$ Reject</th>
<th>$H_A^S$ Reject</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.20</td>
<td>.20</td>
<td>.80</td>
</tr>
<tr>
<td>W</td>
<td>.078</td>
<td>.243</td>
<td>.180</td>
<td>.806</td>
</tr>
<tr>
<td>LR</td>
<td>.057</td>
<td>.198</td>
<td>.180</td>
<td>.758</td>
</tr>
<tr>
<td>LM</td>
<td>.046</td>
<td>.141</td>
<td>.151</td>
<td>.644</td>
</tr>
<tr>
<td>W</td>
<td>.113</td>
<td>.283</td>
<td>.161</td>
<td>.811</td>
</tr>
<tr>
<td>LR</td>
<td>.066</td>
<td>.200</td>
<td>.168</td>
<td>.720</td>
</tr>
<tr>
<td>LM</td>
<td>.042</td>
<td>.110</td>
<td>.127</td>
<td>.507</td>
</tr>
<tr>
<td>W</td>
<td>.214</td>
<td>.410</td>
<td>.140</td>
<td>.838</td>
</tr>
<tr>
<td>LR</td>
<td>.106</td>
<td>.253</td>
<td>.151</td>
<td>.703</td>
</tr>
<tr>
<td>LM</td>
<td>.033</td>
<td>.071</td>
<td>.098</td>
<td>.319</td>
</tr>
</tbody>
</table>

One immediate observation from these aggregated results is that, when the nominal $\chi^2$ critical value is used, $W$ is more likely to reject the null hypothesis than $LR$ and $LR$ is more likely to reject it than $LM$. In fact, the only exception to this in the detailed results of Appendix A can be found in Table (viii) where, for 80 observations and the weak
alternative, \( \Pr(LM > c) = .206 \) while \( \Pr(LR > c) = .205 \), but this difference is almost certainly statistically insignificant. This suggests that the relationship \( W > LR > LM \), similar to that discussed in Chapter 2 for tests of hypotheses about regression coefficients in linear models, might hold on average in this application of the three criteria. Unlike the situation in Chapter 2, however, the inequality relationship is not systematic: an examination of some simulated samples revealed all possible configurations of relative magnitudes of the computed statistics.

As expected, the simulated rejection probabilities are closer to those predicted by asymptotic theory with increasing sample size. Uniform convergence (particularly in the detailed tables in the Appendix) cannot be expected because the exogenous data sets for larger samples were obtained by adding extra new observations to those for smaller samples. On average, asymptotic predictions are closer to the observed behaviour of \( LM \) under the null hypothesis, to \( LR \) for the weak alternative and to \( W \) for the strongly heteroscedastic alternative.

This same phenomenon is also revealed in the power estimates (i.e. estimates of rejection probabilities after the critical values have been adjusted so that Type I error estimates are .05). Overall, the asymptotic power formula would tend to overpredict power, although in Tables (iv), (vi) and (viii) there are individual estimates which exceed the asymptotic prediction for the weak alternative by amounts which are typically statistically insignificant. In the aggregated table above, it would appear that generally \( LR \) is slightly more powerful than \( W \) and that both of these are considerably more powerful than \( LM \) in smaller samples. However an examination of the detailed tables shows individual experiments in which each of \( W, LR \) and \( LM \) dominates the other two procedures. While most of the differences between power estimates are quite small and
hence probably masked by experimental errors, there is a definite suggestion (which is bolstered by using the results from different experiments as collateral information) that there are some exogenous variable sets in which the LM test might be the most powerful and others in which it is the least powerful of the three procedures.

A disturbing feature of the LM test in small samples is apparent from Tables (iii), (vi) and (ix). With 20 observations on the exogenous variable labelled as $X_4$ - realizations of a $U(1,31)$ - as the variance factor, the probability of rejecting a false null hypothesis with the LM criterion can in fact be lower than that of rejecting a true null with the same critical value, even when heteroscedasticity is quite strong.\(^6\) With 40 observations, the anomaly is removed but the LM test still shows low power relative to $W$ and LM tests. In the same situations, asymptotic theory would be a poor guide to the true significance points of both $W$ and LR statistics: for example, in Table (vi) the probability that a $W$ test using the .05 nominal $\chi^2(1)$ critical value would falsely indicate heteroscedasticity in a homoscedastic model with 20 observations is estimated as .388, and for LR the corresponding estimate is .188. These experiments in which asymptotic predictions show poorly as guides to finite sample performance are the situations that were noted in §7.2 as providing the most difficulty with convergence of the estimation algorithm for the full model. Therefore, while there are cases in which the LM test is useless in small samples, application of W and LR

---

\(^6\) A check on the simulated rejection probabilities for the LM test with $X_4$ as the disturbance factor was run using the numerical integration routine discussed in §6.5(ii), suitably modified for heteroscedastic disturbances. The estimates accord closely with the exact probabilities so obtained, allowing for the expected experimental error in the former.
procedures will be difficult and unreliable in the same situations.

There seem to be two interesting questions to be answered: (a) how well, on balance, does asymptotic theory stand up as a predictor of finite-sample behaviour, and (b) what are the attributes of the exogenous observations that account for the, sometimes marked, differences between the three asymptotically equivalent testing criteria?

To examine the first question, the relationships between the sampling estimates $\hat{P}$ and the asymptotic predictions $P^*$ for the fixed nominal critical value are examined for $W$, $LR$ and $LM$ separately, over all 81 experiments. With these 81 observations, regressing $\hat{P}$ on a constant and $P^*$ gives:

$$\hat{P}_W = 0.097 + 0.910P^*$$

$$SSE = 1.130 \quad R^2 = 0.853 \quad DW = 0.98$$

$$\hat{P}_{LR} = 0.035 + 0.867P^*$$

$$SSE = 0.228 \quad R^2 = 0.968 \quad DW = 1.57$$

$$\hat{P}_{LM} = 0.006 + 0.600P^*$$

$$SSE = 1.479 \quad R^2 = 0.708 \quad DW = 1.30$$

Here the figures in parentheses below coefficients are estimated standard errors, $SSE$ is the unexplained residual sum of squares, $R^2$ is the usual coefficient of determination and $DW$ is the Durbin-Watson statistic.

While the latter is only meaningful within the ordering of observations that was used, it is reported as a warning of the unsuitability of the

---

7 All regressions are GLS, allowing for the expected variation in sampling error from null hypothesis to alternative hypothesis situations.
above regressions as "models" of the experimental results. Bearing this
caution in mind, it can be seen that the hypothesis that the coefficient of
P* is unity would be rejected by the usual criteria in every case and,
exception for the LM test, the constant term would be judged to be
significantly different from zero. Overall, it can be seen that the LR
rejection probabilities are best explained by asymptotic predictions,
followed by those of the W criterion with LM behaviour least well
predicted.

On the question of the role played by the exogenous variables, the
first few moments of both X and Z might be expected to capture a large
part of the influence of conditioning on different exogenous data sets.
For this purpose, empirical moments about the sample mean up to the fourth
were used. Relationships of the form

\[ \hat{P} = \gamma_0 P^* + n^{-1}(\gamma_1 + \gamma_2 P^* + \gamma_3 X^2 + \gamma_4 X^3 + \gamma_5 \hat{X}^4 + \gamma_6 \hat{Z}^2 + \gamma_7 \hat{Z}^3 + \gamma_8 \hat{Z}^4) \]

were fitted, treating the three equations for W, LR and LM as a multi­
variate regression. Coefficients \( \gamma_0 \) were found to be severally and
jointly insignificantly different from one [t = .35, t = 1.80, t = 1.87,
and \( \chi^2(3) = 6.32 \) by the W test], so all \( \gamma_0 \) were constrained to be
unity to give \( \hat{P} - P^* \) as the dependent variable.\(^8\) This then yielded
the coefficient estimates and t-ratios which are arrayed in the following
table.

\(^8\) Thus the relationships have appropriate "long-run" properties with
\( P \rightarrow P^* \) as \( n \rightarrow \infty \), when \( \hat{P} \) is interpreted as \( P \) with added
observational error.
<table>
<thead>
<tr>
<th></th>
<th>l</th>
<th>( p^* )</th>
<th>( \bar{x}^2 )</th>
<th>( \bar{x}^3 )</th>
<th>( \bar{x}^4 )</th>
<th>( \bar{z}^2 )</th>
<th>( \bar{z}^3 )</th>
<th>( \bar{z}^4 )</th>
<th>( \text{SSE} )</th>
<th>( R^2 )</th>
<th>DW</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W )</td>
<td>2.35</td>
<td>-3.24</td>
<td>.034</td>
<td>-.0238</td>
<td>.0310</td>
<td>75.44</td>
<td>-183.4</td>
<td>109.5</td>
<td>.249</td>
<td>.825</td>
<td>1.82</td>
</tr>
<tr>
<td>( LR )</td>
<td>.651</td>
<td>-4.35</td>
<td>.028</td>
<td>-.0234</td>
<td>.0396</td>
<td>34.36</td>
<td>-85.50</td>
<td>52.12</td>
<td>.096</td>
<td>.805</td>
<td>1.58</td>
</tr>
<tr>
<td>( LM )</td>
<td>.452</td>
<td>-13.08</td>
<td>.038</td>
<td>-.0250</td>
<td>.0315</td>
<td>-51.09</td>
<td>120.1</td>
<td>-69.66</td>
<td>.670</td>
<td>.778</td>
<td>1.53</td>
</tr>
</tbody>
</table>

\textit{N.B.} Figures in parentheses are "t-ratios".
The intention of this presentation is not to provide a model of the properties of the statistics from which their behaviour in other situations could be predicted (although it is rather suggestive for that usage), but instead to summarize the important influences which explain the discrepancies between $P$ and $P^*$. Observe that the moments of the regressor, $X$, are individually insignificant: they are also jointly insignificant with $\chi^2(9) = 8.90$ for the hypothesis which constrains all of their coefficients to be zeros. The hypothesis that the influence of the $X's$ is the same in each equation is also not rejected by $\chi^2(6) = 3.40$. However, for the variance factor, $Z$, there is a significant difference across the three equations in the response of $(P - P^*)$ to the moments of the $Z$ factor which is used, with $\chi^2(6) = 123.7$ for the hypothesis which constrains coefficients to be equal across equations. Sample size as measured by $1/n$ and strength of heteroscedasticity as measured by $P^*/n$ are also important influences.

These regressions summarize what might also have been deduced from a detailed scrutiny of the tables in Appendix A. In particular, the influence of the variance factor is much stronger than that of the regressors in causing small sample properties to deviate from asymptotic theory predictions. There are some exogenous data sets for which each of $W$, $LR$ and $LM$ has attractive properties, but sometimes discrepancies between asymptotic predictions and finite sample behaviour can be substantial.
APPENDIX A.

Experimental Results*

Table (i)

Regressors: $X_1$ and $X_2$  Variance Factor: $X_2$

<table>
<thead>
<tr>
<th>Nominal</th>
<th>$H_0$</th>
<th>$H_A(W)$ Reject Power</th>
<th>$H_A(S)$ Reject Power</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.20 .20</td>
<td>.80 .80</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>.086</td>
<td>.270 .194</td>
<td>.820 .756</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>.054</td>
<td>.206 .197</td>
<td>.757 .748</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>.035</td>
<td>.132 .168</td>
<td>.615 .680</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>.116</td>
<td>.314 .163</td>
<td>.824 .674</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>.064</td>
<td>.207 .183</td>
<td>.719 .681</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>.040</td>
<td>.103 .123</td>
<td>.498 .526</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>.175</td>
<td>.367 .158</td>
<td>.821 .591</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>.084</td>
<td>.232 .169</td>
<td>.683 .598</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>.032</td>
<td>.103 .139</td>
<td>.401 .494</td>
<td></td>
</tr>
</tbody>
</table>

Table (ii)

Regressors: $X_1$ and $X_2$  Variance Factor: $X_3$

<table>
<thead>
<tr>
<th>Nominal</th>
<th>$H_0$</th>
<th>$H_A(W)$ Reject Power</th>
<th>$H_A(S)$ Reject Power</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.20 .20</td>
<td>.80 .80</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>.077</td>
<td>.215 .159</td>
<td>.788 .723</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>.060</td>
<td>.198 .169</td>
<td>.772 .737</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>.051</td>
<td>.189 .187</td>
<td>.749 .747</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>.094</td>
<td>.233 .158</td>
<td>.786 .699</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>.055</td>
<td>.188 .181</td>
<td>.738 .728</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>.036</td>
<td>.176 .199</td>
<td>.686 .715</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>.164</td>
<td>.374 .162</td>
<td>.839 .607</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>.081</td>
<td>.245 .169</td>
<td>.737 .650</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>.026</td>
<td>.117 .170</td>
<td>.508 .603</td>
<td></td>
</tr>
</tbody>
</table>

* These tables are described at the beginning of §7.5 on pp.172-173 above.
Appendix A/Continued

Table (iii)
Regressors: $X_1$ and $X_2$  Variance Factor: $X_4$

<table>
<thead>
<tr>
<th>Nominal</th>
<th>$H_0$</th>
<th>$H_A(W)$</th>
<th>$H_A(S)$</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Reject Power</td>
<td>Reject Power</td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td>.20</td>
<td>.20</td>
<td>.80</td>
<td>.80</td>
</tr>
<tr>
<td>$W$</td>
<td>.077</td>
<td>.246</td>
<td>.178</td>
<td>.829</td>
</tr>
<tr>
<td>$LR$</td>
<td>.060</td>
<td>.186</td>
<td>.167</td>
<td>.765</td>
</tr>
<tr>
<td>$LM$</td>
<td>.057</td>
<td>.090</td>
<td>.078</td>
<td>.580</td>
</tr>
<tr>
<td>$W$</td>
<td>.110</td>
<td>.310</td>
<td>.185</td>
<td>.833</td>
</tr>
<tr>
<td>$LR$</td>
<td>.065</td>
<td>.200</td>
<td>.169</td>
<td>.694</td>
</tr>
<tr>
<td>$LM$</td>
<td>.044</td>
<td>.047</td>
<td>.068</td>
<td>.322</td>
</tr>
<tr>
<td>$W$</td>
<td>.247</td>
<td>.468</td>
<td>.132</td>
<td>.868</td>
</tr>
<tr>
<td>$LR$</td>
<td>.128</td>
<td>.276</td>
<td>.125</td>
<td>.688</td>
</tr>
<tr>
<td>$LM$</td>
<td>.033</td>
<td>.004</td>
<td>.018</td>
<td>.021</td>
</tr>
</tbody>
</table>

Table (iv)
Regressors: $X_1$ and $X_3$  Variance Factor: $X_2$

<table>
<thead>
<tr>
<th>Nominal</th>
<th>$H_0$</th>
<th>$H_A(W)$</th>
<th>$H_A(S)$</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Reject Power</td>
<td>Reject Power</td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td>.20</td>
<td>.20</td>
<td>.80</td>
<td>.80</td>
</tr>
<tr>
<td>$W$</td>
<td>.078</td>
<td>.267</td>
<td>.204</td>
<td>.823</td>
</tr>
<tr>
<td>$LR$</td>
<td>.059</td>
<td>.217</td>
<td>.190</td>
<td>.776</td>
</tr>
<tr>
<td>$LM$</td>
<td>.055</td>
<td>.154</td>
<td>.139</td>
<td>.669</td>
</tr>
<tr>
<td>$W$</td>
<td>.099</td>
<td>.269</td>
<td>.159</td>
<td>.820</td>
</tr>
<tr>
<td>$LR$</td>
<td>.063</td>
<td>.199</td>
<td>.169</td>
<td>.758</td>
</tr>
<tr>
<td>$LM$</td>
<td>.049</td>
<td>.122</td>
<td>.124</td>
<td>.580</td>
</tr>
<tr>
<td>$W$</td>
<td>.141</td>
<td>.303</td>
<td>.162</td>
<td>.789</td>
</tr>
<tr>
<td>$LR$</td>
<td>.069</td>
<td>.200</td>
<td>.159</td>
<td>.691</td>
</tr>
<tr>
<td>$LM$</td>
<td>.034</td>
<td>.098</td>
<td>.123</td>
<td>.433</td>
</tr>
</tbody>
</table>
Appendix A/Continued

Table (v)

Regressors: $X_1$ and $X_3$, Variance Factor: $X_3$

<table>
<thead>
<tr>
<th>Nominal</th>
<th>$H_0$</th>
<th>$H_{A(W)}$</th>
<th>$H_{A(S)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.20</td>
<td>.20</td>
</tr>
<tr>
<td>W</td>
<td>.088</td>
<td>.181</td>
<td>.118</td>
</tr>
<tr>
<td>LR</td>
<td>.062</td>
<td>.162</td>
<td>.135</td>
</tr>
<tr>
<td>LM</td>
<td>.043</td>
<td>.151</td>
<td>.165</td>
</tr>
<tr>
<td>W</td>
<td>.130</td>
<td>.203</td>
<td>.080</td>
</tr>
<tr>
<td>LR</td>
<td>.080</td>
<td>.149</td>
<td>.098</td>
</tr>
<tr>
<td>LM</td>
<td>.037</td>
<td>.103</td>
<td>.121</td>
</tr>
<tr>
<td>W</td>
<td>.114</td>
<td>.254</td>
<td>.147</td>
</tr>
<tr>
<td>LR</td>
<td>.069</td>
<td>.175</td>
<td>.147</td>
</tr>
<tr>
<td>LM</td>
<td>.038</td>
<td>.107</td>
<td>.129</td>
</tr>
</tbody>
</table>

Table (vi)

Regressors: $X_1$ and $X_3$, Variance Factor: $X_4$

<table>
<thead>
<tr>
<th>Nominal</th>
<th>$H_0$</th>
<th>$H_{A(W)}$</th>
<th>$H_{A(S)}$</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.20</td>
<td>.20</td>
<td>.80</td>
</tr>
<tr>
<td>W</td>
<td>.071</td>
<td>.272</td>
<td>.219</td>
<td>.821</td>
</tr>
<tr>
<td>LR</td>
<td>.048</td>
<td>.208</td>
<td>.216</td>
<td>.759</td>
</tr>
<tr>
<td>LM</td>
<td>.037</td>
<td>.110</td>
<td>.137</td>
<td>.558</td>
</tr>
<tr>
<td>W</td>
<td>.144</td>
<td>.347</td>
<td>.171</td>
<td>.854</td>
</tr>
<tr>
<td>LR</td>
<td>.078</td>
<td>.233</td>
<td>.173</td>
<td>.736</td>
</tr>
<tr>
<td>LM</td>
<td>.044</td>
<td>.072</td>
<td>.089</td>
<td>.372</td>
</tr>
<tr>
<td>W</td>
<td>.388</td>
<td>.615</td>
<td>.111</td>
<td>.904</td>
</tr>
<tr>
<td>LR</td>
<td>.188</td>
<td>.378</td>
<td>.128</td>
<td>.713</td>
</tr>
<tr>
<td>LM</td>
<td>.032</td>
<td>.005</td>
<td>.026</td>
<td>.013</td>
</tr>
</tbody>
</table>
Appendix A/Continued

Table (vii)
Regressors: $X_1$ and $X_4$  Variance Factor: $X_2$

<table>
<thead>
<tr>
<th>Nominal</th>
<th>$H_0$ Reject Power</th>
<th>$H_A(W)$ Reject Power</th>
<th>$H_A(S)$ Reject Power</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.20</td>
<td>.20</td>
<td>.80</td>
</tr>
<tr>
<td>W</td>
<td>.084</td>
<td>.253</td>
<td>.175</td>
<td>.830</td>
</tr>
<tr>
<td>LR</td>
<td>.058</td>
<td>.201</td>
<td>.171</td>
<td>.766</td>
</tr>
<tr>
<td>LM</td>
<td>.045</td>
<td>.132</td>
<td>.144</td>
<td>.653</td>
</tr>
<tr>
<td>W</td>
<td>.098</td>
<td>.288</td>
<td>.180</td>
<td>.832</td>
</tr>
<tr>
<td>LR</td>
<td>.062</td>
<td>.203</td>
<td>.176</td>
<td>.740</td>
</tr>
<tr>
<td>LM</td>
<td>.045</td>
<td>.120</td>
<td>.127</td>
<td>.551</td>
</tr>
<tr>
<td>W</td>
<td>.180</td>
<td>.362</td>
<td>.144</td>
<td>.835</td>
</tr>
<tr>
<td>LR</td>
<td>.084</td>
<td>.241</td>
<td>.171</td>
<td>.728</td>
</tr>
<tr>
<td>LM</td>
<td>.036</td>
<td>.112</td>
<td>.137</td>
<td>.488</td>
</tr>
</tbody>
</table>

Table (viii)
Regressors: $X_1$ and $X_4$  Variance Factor: $X_3$

<table>
<thead>
<tr>
<th>Nominal</th>
<th>$H_0$ Reject Power</th>
<th>$H_A(W)$ Reject Power</th>
<th>$H_A(S)$ Reject Power</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.20</td>
<td>.20</td>
<td>.80</td>
</tr>
<tr>
<td>W</td>
<td>.066</td>
<td>.223</td>
<td>.174</td>
<td>.798</td>
</tr>
<tr>
<td>LR</td>
<td>.050</td>
<td>.205</td>
<td>.207</td>
<td>.781</td>
</tr>
<tr>
<td>LM</td>
<td>.041</td>
<td>.206</td>
<td>.231</td>
<td>.744</td>
</tr>
<tr>
<td>W</td>
<td>.090</td>
<td>.241</td>
<td>.165</td>
<td>.799</td>
</tr>
<tr>
<td>LR</td>
<td>.057</td>
<td>.203</td>
<td>.190</td>
<td>.756</td>
</tr>
<tr>
<td>LM</td>
<td>.040</td>
<td>.185</td>
<td>.212</td>
<td>.692</td>
</tr>
<tr>
<td>W</td>
<td>.185</td>
<td>.342</td>
<td>.122</td>
<td>.831</td>
</tr>
<tr>
<td>LR</td>
<td>.095</td>
<td>.210</td>
<td>.141</td>
<td>.713</td>
</tr>
<tr>
<td>LM</td>
<td>.038</td>
<td>.086</td>
<td>.117</td>
<td>.479</td>
</tr>
</tbody>
</table>
Appendix A/Continued

Table (ix)

Regressors: $X_1$ and $X_4$ Variance Factor: $X_4$

<table>
<thead>
<tr>
<th>Nominal</th>
<th>$H_0$</th>
<th>$H_A(W)$ Reject Power</th>
<th>$H_A(S)$ Reject Power</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.05</td>
<td>.20</td>
<td>.20</td>
<td>.80</td>
</tr>
<tr>
<td>W</td>
<td>.078</td>
<td>.260</td>
<td>.199</td>
<td>.819</td>
</tr>
<tr>
<td>LR</td>
<td>.061</td>
<td>.198</td>
<td>.171</td>
<td>.758</td>
</tr>
<tr>
<td>LM</td>
<td>.049</td>
<td>.106</td>
<td>.108</td>
<td>.580</td>
</tr>
<tr>
<td>W</td>
<td>.133</td>
<td>.341</td>
<td>.189</td>
<td>.841</td>
</tr>
<tr>
<td>LR</td>
<td>.072</td>
<td>.220</td>
<td>.175</td>
<td>.698</td>
</tr>
<tr>
<td>LM</td>
<td>.041</td>
<td>.061</td>
<td>.077</td>
<td>.337</td>
</tr>
<tr>
<td>W</td>
<td>.328</td>
<td>.603</td>
<td>.125</td>
<td>.893</td>
</tr>
<tr>
<td>LR</td>
<td>.152</td>
<td>.316</td>
<td>.151</td>
<td>.700</td>
</tr>
<tr>
<td>LM</td>
<td>.029</td>
<td>.004</td>
<td>.021</td>
<td>.014</td>
</tr>
</tbody>
</table>
### APPENDIX B.  
**Exogenous Variables Data**

<table>
<thead>
<tr>
<th>OBS</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>49.51</td>
<td>48.96</td>
<td>25.67</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>42.40</td>
<td>14.07</td>
<td>18.79</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>58.58</td>
<td>19.31</td>
<td>23.43</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>55.30</td>
<td>61.63</td>
<td>1.16</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>57.74</td>
<td>50.31</td>
<td>20.95</td>
</tr>
<tr>
<td>6</td>
<td>1.00</td>
<td>53.42</td>
<td>10.31</td>
<td>11.82</td>
</tr>
<tr>
<td>7</td>
<td>1.00</td>
<td>55.02</td>
<td>11.44</td>
<td>24.29</td>
</tr>
<tr>
<td>8</td>
<td>1.00</td>
<td>46.61</td>
<td>19.19</td>
<td>26.76</td>
</tr>
<tr>
<td>9</td>
<td>1.00</td>
<td>34.32</td>
<td>10.01</td>
<td>8.33</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>41.48</td>
<td>19.02</td>
<td>26.12</td>
</tr>
<tr>
<td>11</td>
<td>1.00</td>
<td>45.68</td>
<td>49.60</td>
<td>9.70</td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>62.53</td>
<td>23.38</td>
<td>12.59</td>
</tr>
<tr>
<td>13</td>
<td>1.00</td>
<td>68.05</td>
<td>3.82</td>
<td>22.40</td>
</tr>
<tr>
<td>14</td>
<td>1.00</td>
<td>58.27</td>
<td>11.56</td>
<td>22.00</td>
</tr>
<tr>
<td>15</td>
<td>1.00</td>
<td>49.40</td>
<td>8.25</td>
<td>26.25</td>
</tr>
<tr>
<td>16</td>
<td>1.00</td>
<td>43.05</td>
<td>23.76</td>
<td>29.77</td>
</tr>
<tr>
<td>17</td>
<td>1.00</td>
<td>66.63</td>
<td>53.72</td>
<td>14.39</td>
</tr>
<tr>
<td>18</td>
<td>1.00</td>
<td>50.94</td>
<td>18.79</td>
<td>18.37</td>
</tr>
<tr>
<td>19</td>
<td>1.00</td>
<td>55.66</td>
<td>55.76</td>
<td>13.93</td>
</tr>
<tr>
<td>20</td>
<td>1.00</td>
<td>47.40</td>
<td>3.98</td>
<td>8.19</td>
</tr>
<tr>
<td>21</td>
<td>1.00</td>
<td>27.97</td>
<td>38.20</td>
<td>28.19</td>
</tr>
<tr>
<td>22</td>
<td>1.00</td>
<td>25.82</td>
<td>56.32</td>
<td>15.67</td>
</tr>
<tr>
<td>23</td>
<td>1.00</td>
<td>60.06</td>
<td>125.06</td>
<td>14.38</td>
</tr>
<tr>
<td>24</td>
<td>1.00</td>
<td>70.56</td>
<td>47.31</td>
<td>16.41</td>
</tr>
<tr>
<td>25</td>
<td>1.00</td>
<td>50.33</td>
<td>14.99</td>
<td>28.64</td>
</tr>
<tr>
<td>26</td>
<td>1.00</td>
<td>49.24</td>
<td>58.10</td>
<td>11.36</td>
</tr>
<tr>
<td>27</td>
<td>1.00</td>
<td>52.31</td>
<td>23.32</td>
<td>12.40</td>
</tr>
<tr>
<td>28</td>
<td>1.00</td>
<td>57.47</td>
<td>10.65</td>
<td>5.02</td>
</tr>
<tr>
<td>29</td>
<td>1.00</td>
<td>46.41</td>
<td>266.24</td>
<td>4.90</td>
</tr>
<tr>
<td>30</td>
<td>1.00</td>
<td>52.28</td>
<td>7.69</td>
<td>21.51</td>
</tr>
<tr>
<td>31</td>
<td>1.00</td>
<td>37.31</td>
<td>13.38</td>
<td>8.73</td>
</tr>
<tr>
<td>32</td>
<td>1.00</td>
<td>44.62</td>
<td>8.24</td>
<td>14.99</td>
</tr>
<tr>
<td>33</td>
<td>1.00</td>
<td>41.25</td>
<td>10.69</td>
<td>27.71</td>
</tr>
<tr>
<td>34</td>
<td>1.00</td>
<td>48.09</td>
<td>18.40</td>
<td>20.87</td>
</tr>
<tr>
<td>35</td>
<td>1.00</td>
<td>56.24</td>
<td>12.40</td>
<td>30.33</td>
</tr>
<tr>
<td>36</td>
<td>1.00</td>
<td>40.08</td>
<td>112.91</td>
<td>11.24</td>
</tr>
<tr>
<td>37</td>
<td>1.00</td>
<td>42.29</td>
<td>4.23</td>
<td>9.74</td>
</tr>
<tr>
<td>38</td>
<td>1.00</td>
<td>34.67</td>
<td>26.95</td>
<td>24.37</td>
</tr>
<tr>
<td>39</td>
<td>1.00</td>
<td>49.56</td>
<td>8.76</td>
<td>20.94</td>
</tr>
<tr>
<td>40</td>
<td>1.00</td>
<td>38.13</td>
<td>26.57</td>
<td>3.98</td>
</tr>
</tbody>
</table>

*cont.*
### Appendix B/Continued

<table>
<thead>
<tr>
<th>OBS</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>1.00</td>
<td>55.48</td>
<td>38.71</td>
<td>23.86</td>
</tr>
<tr>
<td>42</td>
<td>1.00</td>
<td>44.69</td>
<td>62.31</td>
<td>23.23</td>
</tr>
<tr>
<td>43</td>
<td>1.00</td>
<td>56.15</td>
<td>22.30</td>
<td>22.16</td>
</tr>
<tr>
<td>44</td>
<td>1.00</td>
<td>49.99</td>
<td>7.29</td>
<td>8.98</td>
</tr>
<tr>
<td>45</td>
<td>1.00</td>
<td>53.53</td>
<td>147.25</td>
<td>24.51</td>
</tr>
<tr>
<td>46</td>
<td>1.00</td>
<td>58.06</td>
<td>17.29</td>
<td>15.98</td>
</tr>
<tr>
<td>47</td>
<td>1.00</td>
<td>37.58</td>
<td>34.14</td>
<td>23.64</td>
</tr>
<tr>
<td>48</td>
<td>1.00</td>
<td>48.12</td>
<td>48.49</td>
<td>22.10</td>
</tr>
<tr>
<td>49</td>
<td>1.00</td>
<td>41.66</td>
<td>11.33</td>
<td>2.44</td>
</tr>
<tr>
<td>50</td>
<td>1.00</td>
<td>63.20</td>
<td>25.38</td>
<td>13.20</td>
</tr>
<tr>
<td>51</td>
<td>1.00</td>
<td>50.36</td>
<td>15.10</td>
<td>20.04</td>
</tr>
<tr>
<td>52</td>
<td>1.00</td>
<td>40.35</td>
<td>209.90</td>
<td>19.67</td>
</tr>
<tr>
<td>53</td>
<td>1.00</td>
<td>53.99</td>
<td>20.38</td>
<td>6.30</td>
</tr>
<tr>
<td>54</td>
<td>1.00</td>
<td>47.34</td>
<td>21.44</td>
<td>13.62</td>
</tr>
<tr>
<td>55</td>
<td>1.00</td>
<td>43.58</td>
<td>42.61</td>
<td>6.86</td>
</tr>
<tr>
<td>56</td>
<td>1.00</td>
<td>75.37</td>
<td>6.22</td>
<td>8.34</td>
</tr>
<tr>
<td>57</td>
<td>1.00</td>
<td>49.29</td>
<td>7.56</td>
<td>28.10</td>
</tr>
<tr>
<td>58</td>
<td>1.00</td>
<td>42.70</td>
<td>5.75</td>
<td>6.30</td>
</tr>
<tr>
<td>59</td>
<td>1.00</td>
<td>59.56</td>
<td>28.63</td>
<td>24.87</td>
</tr>
<tr>
<td>60</td>
<td>1.00</td>
<td>47.59</td>
<td>15.26</td>
<td>22.54</td>
</tr>
<tr>
<td>61</td>
<td>1.00</td>
<td>67.53</td>
<td>8.39</td>
<td>16.31</td>
</tr>
<tr>
<td>62</td>
<td>1.00</td>
<td>44.92</td>
<td>25.60</td>
<td>8.22</td>
</tr>
<tr>
<td>63</td>
<td>1.00</td>
<td>55.60</td>
<td>34.61</td>
<td>15.10</td>
</tr>
<tr>
<td>64</td>
<td>1.00</td>
<td>60.97</td>
<td>3.02</td>
<td>26.98</td>
</tr>
<tr>
<td>65</td>
<td>1.00</td>
<td>33.55</td>
<td>10.20</td>
<td>22.13</td>
</tr>
<tr>
<td>66</td>
<td>1.00</td>
<td>51.78</td>
<td>11.83</td>
<td>19.14</td>
</tr>
<tr>
<td>67</td>
<td>1.00</td>
<td>51.34</td>
<td>7.17</td>
<td>3.32</td>
</tr>
<tr>
<td>68</td>
<td>1.00</td>
<td>72.22</td>
<td>12.19</td>
<td>22.51</td>
</tr>
<tr>
<td>69</td>
<td>1.00</td>
<td>52.88</td>
<td>23.35</td>
<td>10.22</td>
</tr>
<tr>
<td>70</td>
<td>1.00</td>
<td>51.01</td>
<td>25.92</td>
<td>20.97</td>
</tr>
<tr>
<td>71</td>
<td>1.00</td>
<td>45.40</td>
<td>8.46</td>
<td>23.01</td>
</tr>
<tr>
<td>72</td>
<td>1.00</td>
<td>47.69</td>
<td>50.66</td>
<td>8.71</td>
</tr>
<tr>
<td>73</td>
<td>1.00</td>
<td>49.92</td>
<td>11.95</td>
<td>2.70</td>
</tr>
<tr>
<td>74</td>
<td>1.00</td>
<td>36.87</td>
<td>19.77</td>
<td>22.57</td>
</tr>
<tr>
<td>75</td>
<td>1.00</td>
<td>52.34</td>
<td>38.18</td>
<td>3.46</td>
</tr>
<tr>
<td>76</td>
<td>1.00</td>
<td>61.92</td>
<td>7.58</td>
<td>11.01</td>
</tr>
<tr>
<td>77</td>
<td>1.00</td>
<td>51.54</td>
<td>66.41</td>
<td>12.12</td>
</tr>
<tr>
<td>78</td>
<td>1.00</td>
<td>23.02</td>
<td>36.51</td>
<td>5.54</td>
</tr>
<tr>
<td>79</td>
<td>1.00</td>
<td>41.88</td>
<td>35.29</td>
<td>18.01</td>
</tr>
<tr>
<td>80</td>
<td>1.00</td>
<td>49.08</td>
<td>126.19</td>
<td>22.65</td>
</tr>
</tbody>
</table>
APPENDIX C.

Invariance of the Test Statistics to Certain Parameter Values

Consider the normal linear regression model with \( n \) observations, having exogenous regressors and nonscalar covariance matrix:

\[
y = X\beta + u
\]

\[
u \sim N(0, \sigma^2\Omega)
\]

\[
\Omega = \Omega(\theta, Z) .
\]

Apart from the scale factor \( \sigma^2 \), the covariance matrix depends upon a vector \( \theta \) of parameters which are unrelated to the \( \beta \) regression coefficients and also possibly depends upon some exogenous observations contained in \( Z \). It is assumed that the observations in \( X \) and \( Z \) and the parameterization of \( \Omega \) by \( \theta \) enable the true parameter values \( \beta_o, \sigma^2_o \) and \( \theta_o \) to be identified.

**THEOREM:** The exact finite-sample distributions, conditional upon \( (X, Z) \), of

1. the maximum likelihood estimator of \( \theta_o \), and
2. the \( W \), LR and LM statistics for testing hypotheses about \( \theta_o \),

do not depend upon the parameter values \( \beta_o \) and \( \sigma^2_o \).

**Proof:** For the above model, the log-likelihood is

\[
\ell(\beta, \sigma^2, \theta, y) = c_1 - \frac{1}{2} \log |\Omega| - \frac{1}{2} n \log \sigma^2 - \frac{1}{2} \sigma^{-2}(y-X\beta)'\Omega^{-1}(y-X\beta)
\]

where \( c_1 = -\frac{1}{2} n \log (2\pi) \) and \( \Omega \) is the function of \( \theta \) and \( Z \) given in (3).\(^1\) Maximization of (4) with respect to \( (\beta, \sigma^2, \theta) \) requires,

---

\(^1\) The obvious dependence of \( \ell(\cdot) \) upon \( (X, Z) \) is not shown explicitly for reasons of clarity, since these remain constant throughout.
inter alia,

\[
\frac{\partial \ell}{\partial \beta} = 0 \implies \beta = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y \tag{5}
\]

\[
\frac{\partial \ell}{\partial \sigma^2} = 0 \implies \sigma^2 = n^{-1} (y - X\beta)' \Omega^{-1} (y - X\beta) . \tag{6}
\]

An alternative representation of the maximum likelihood problem can be given in terms of the concentrated (condensed or reduced) log-likelihood function. This is the function of \( \theta \) alone, given the data, obtained by substituting in (4) for \( \beta \) and \( \sigma^2 \) from (5) and (6):

\[
\xi_c(\theta, y) = c_2 - \frac{1}{2} \log |\Omega| - \frac{1}{2} n \log[y' \Omega^{-1} y - y' \Omega^{-1} X (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y]
\]

where \( c_2 = c_1 - \frac{1}{2} n (1 - \log n) \). By construction, if \((\hat{\beta}, \hat{\sigma}^2, \hat{\theta})\) jointly maximize (4) then \( \hat{\theta} \) will also maximize (7), and \((\hat{\beta}, \hat{\sigma}^2)\) will be given from (5) and (6) by substitution of \( \hat{\theta} \). Also

\[
\xi(\hat{\beta}, \hat{\sigma}^2, \hat{\theta}, y) = \xi_c(\hat{\theta}, y) .
\]

Now consider (7) not in its usual role as a function of the parameters as variable arguments for a given realization of the random vector \( y \) which constitutes the data, but instead as a stochastic function of the random variables \( y \). The elements of \( \theta \) which enter \( \Omega \) in (7) are still to be interpreted as variable arguments of \( \xi_c(\theta, y) \). Since \( y \) is a vector of random variables with joint distribution function depending parametrically upon \((\beta_0, \sigma_0^2, \theta_0)\), the distribution of the maximizing values \((\hat{\beta}, \hat{\sigma}^2, \hat{\theta})\) and of the supremum \( \xi(\hat{\beta}, \hat{\sigma}^2, \hat{\theta}, y) \) will generally depend upon the same parameters. However, from (1),

\[
y' \Omega^{-1} y - y' \Omega^{-1} X (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y = u' \Omega^{-1} u - u' \Omega^{-1} X (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} u
\]

so that \( \xi_c(\theta, y) = \xi_c(\theta, u) \). Because \( \beta_0 \) does not enter the distribution
function of \( u \), the distribution of \( \ell_c(\theta, y) \) cannot depend upon \( \beta_0 \). Neither can \( \beta_0 \) affect the location of the maximum of \( \ell_c(\theta, y) \) with respect to the variable argument \( \theta \), nor that of the corresponding supremum of \( \ell_c(\theta, y) \).

To demonstrate the invariance of the distribution of \( \hat{\theta} \) to the parameter \( \sigma^2_o \), the same concepts are used but the setting is a little more complicated. Note that

\[
\ell_c(\theta, y) = \ell_c \left( \theta, \frac{u}{\sigma_0} \right) - \frac{1}{2} n \log \sigma^2_o
\]

so that \( [\ell_c(\theta, y) + \frac{1}{2} n \log \sigma^2_o] \) is invariant to \((\beta_0, \sigma^2_o)\). For any realization of the random variables in the model, the value of \( \theta \) which maximizes \( [\ell_c(\theta, y) + \frac{1}{2} n \log \sigma^2_o] \) also maximizes \( \ell_c(\theta, y) \) so that \( \hat{\theta} \) and \( \tilde{\theta} \) have distributions independent of both \( \beta_0 \) and \( \sigma^2_o \).

Consider now the three principles for testing the hypothesis \( \phi(\theta_o) = 0 \). Provided \( \beta_0 \) and \( \sigma^2_o \) do not enter the hypothesized constraints on \( \theta_o \), the above reasoning will apply equally when maximization is restricted to that region of the parameter space of \( \theta \) in which \( \phi(\theta) = 0 \). Denote the restricted estimates by \((\hat{\theta}, \hat{\sigma}^2, \tilde{\theta})\). Restricted estimation requires maximization with respect to \( \theta \) and \( \lambda \) of the Lagrangian function,

\[
\psi(\theta, \lambda, y) = \ell_c(\theta, y) + \lambda' \phi(\theta)
\]

where the conditions (5) and (6) will remain unchanged to give \((\hat{\theta}, \hat{\sigma}^2)\).

---

2 This may be easier to conceptualize within the framework of a simulation experiment. Suppose that the following steps are taken: (i) specify \((\sigma^2_o, \theta_o)\) and generate \( u \) according to the probability law in (2), then (ii) specify \( \beta_0 \) and generate \( y \) as in (1). The concentrated log-likelihood formed from \( y \) after step (ii) could equally well be formed from \( u \) after step (i), with step (ii) being omitted entirely. Thus the same objective function with the same maximizing value \( \hat{\theta} \) and the same supremum over \( \theta \) can be obtained without even specifying a value for \( \beta_0 \).
directly from $\hat{\theta}$. Note that

$$\psi(\hat{\theta}, \hat{\lambda}, y) = \ell_c(\hat{\theta}, y)$$

so that the distributions of $(\hat{\theta}, \hat{\lambda})$ and $[\ell_c(\hat{\theta}, y) + \frac{1}{2} n \log \sigma_o^2]$ cannot depend upon $(\beta_o, \sigma_o^2)$, and the same applies to the relationship between constrained and unconstrained estimation problems.

The likelihood ratio criterion for testing the hypothesis is

$$LR = 2[\ell(\hat{\beta}, \hat{\sigma}^2, \hat{\theta}, y) - \ell(\hat{\theta}, \hat{\sigma}^2, \hat{\theta}, y)]$$

$$= 2[\ell_c(\hat{\theta}, y) - \ell_c(\hat{\theta}, y)]$$

$$= 2[\{\ell_c(\hat{\theta}, y) + \frac{1}{2} n \log \sigma_o^2\} - \{\ell_c(\hat{\theta}, y) + \frac{1}{2} n \log \sigma_o^2\}]$$

which is the difference between two random variables for which the joint distribution function does not depend upon $(\beta_o, \sigma_o^2)$.

To form the $W$ and $LR$ statistics, the information matrix is required, and this is for $(\beta, \sigma^2, \theta)$:

$$I = \begin{bmatrix} (X'\Omega^{-1}X)^{-1} & 0 & 0 \\ 0 & \frac{1}{2} n \sigma^{-4} & \frac{1}{2} \sigma^{-2} (\text{vec } \Omega^{-1})'A \\ 0 & \frac{1}{2} \sigma^{-2} A' (\text{vec } \Omega^{-1}) & \frac{1}{2} A' (\Omega^{-1} \theta \Omega^{-1})'A \end{bmatrix}$$

where $A = (\theta \text{ vec } \Omega/\theta \theta)'$. Denoting $F = F(\theta) = \partial \phi/\partial \theta(\theta)$ and

$$I^{\theta\theta} = [\frac{1}{2} A' ((\Omega^{-1} \theta \Omega^{-1}) - n^{-1} (\text{vec } \Omega^{-1}) (\text{vec } \Omega^{-1})')A]^{-1}$$

which are functions of $\theta$ parameters alone, the $W$ statistic is

---

3 The present model is slightly different from that in §3.5 because of the explicit factorization of the covariance matrix as $\sigma^2\Omega$. 

$W = [\phi(\hat{\theta})]'(\hat{F}'\hat{I}^{\theta\theta}\hat{F})^{-1}[\phi(\hat{\theta})]$ 

where $\hat{F} = F(\theta)$ and $\hat{I}^{\theta\theta} = I^{\theta\theta}(\theta)$. Since the distribution of $W$ depends on the parameters only through $\hat{\theta}$, it too must be invariant to $(\beta_0, \sigma^2_0)$.

The LM statistic is

$$LM = \lambda'\hat{F}'\hat{I}^{\theta\theta}\hat{F}\lambda$$

for which the distribution depends on the parameters only through the joint distribution of $(\hat{\beta}, \hat{\lambda})$, so that the same conclusions will apply.

As a corollary to the above,

$$\left(\frac{\hat{\beta} - \beta_0}{\sigma_0}\right)^2 = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}\left(\frac{u}{\sigma_0}\right)$$

$$\left(\frac{\hat{\sigma}}{\sigma_0}\right)^2 = n^{-1}\left[\left(\frac{u}{\sigma_0}\right)'\hat{\Omega}^{-1}\left(\frac{u}{\sigma_0}\right) - \left(\frac{u}{\sigma_0}\right)'\hat{\Omega}^{-1}X(X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}\left(\frac{u}{\sigma_0}\right)\right]$$

where $\hat{\Omega} = \Omega(\hat{\theta}, Z)$, will have distributions which depend upon the parameters only through the joint distribution of $(\hat{\theta}, (u/\sigma_0))$ and hence do not depend upon $\beta_0$ or $\sigma^2_0$. Similar conclusions apply for restricted estimators and, it would appear, for many situations in which two-step estimators rather than maximum likelihood ones are of interest.

One obvious application for these results is in the design of Monte Carlo simulation experiments. If interest centres upon estimators or test statistics for various sorts of nonscalar covariance matrices in the normal linear model (related to autocorrelation, heteroscedasticity, error components etc.), then it is pointless to vary those parameters to which the estimators or tests are invariant. If the procedures to be examined have the same properties for all points in the parameter space of some of the parameters, then there is nothing to be gained in simulation studies by designing over a range of values of these parameters.
Equally, nothing is gained by choosing "realistic" values for these parameters. In some situations, such as those represented in (8) and (9), the effect of varying certain parameters can be predicted exactly so that again only subspaces of the parameter space need be examined in the experiments.
CHAPTER 8
CONCLUSIONS

Of the three asymptotic principles for testing hypotheses, Wald (W), likelihood ratio (LR) and Lagrange multiplier (LM), the first two are frequently employed in econometrics but the third is less commonly used. However, a number of procedures that are familiar to the econometrician do correspond to LM tests; this interpretation facilitates an understanding of these procedures and their extension to different situations. Particularly in the context of diagnosing misspecification errors, the LM principles provides a useful framework for developing tests that are asymptotically equivalent to, but less demanding computationally than, the corresponding W and LR tests.

The score test form of the LM criterion has been emphasized throughout, because this formulation has been found to be more straightforward to apply in most situations than working explicitly with the Lagrange multipliers that arise in constrained likelihood maximization. Connections have been drawn between the LM testing procedure and similar methods, such as the C(α) statistic and various criteria based upon estimating the parameters of the model in two stages. These connections serve to relate what may appear to be diverse approaches in the statistics and econometrics literature, and to provide alternative computational schemes for obtaining the LM statistic or a close approximation to it.

Applications of the LM testing principle have demonstrated its usefulness in many econometric situations. The LM criterion is often much easier to compute than the W or LR statistics in the same situation; frequently a regression interpretation allows standard computer programs to be used for calculating a misspecification diagnostic from
the residuals of the fitted model. While applications to the traditional econometric misspecification areas of autocorrelation and heteroscedasticity have been emphasized, the general approach allows applied researchers to construct tests for the particular specifications in which they are interested.

The LM test has considerable optimality properties in asymptotic theory but the evidence on finite-sample performance that has been obtained in a simulation study of one application is less encouraging. Future research could be most profitably directed toward analysis of, and discrimination between, hypothesis testing procedures in practical situations with finite samples.
REFERENCES


Sargan, J.D. (1975), "A suggested technique for computing approximations to Wald criteria with application to testing dynamic specifications", Discussion Paper A2, London School of Economics.


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.

