INSTITUTE FOR ECONOMIC RESEARCH

QUEEN'S UNIVERSITY

Kingston, Ontario, Canada   K7L 3N6

PRINCIPLES AND METHODS IN THE TESTING OF
ALTERNATIVE MODELS

Gordon Fisher and Michael McAleer

Department of Economics
Queen's University

Discussion Paper No. 400

August, 1980

Paper to the North American Meeting of the Econometric Society,
Denver, Colorado, September 5-7, 1980.

CONTENTS

ABSTRACT

This paper seeks to distinguish the principles upon which testing of statistical hypotheses may be based and the practical methods which these principles generate. Six examples are given for the case of nested hypotheses as illustrations. In particular, Seber's (1964) conclusion that the Wald, Lagrange Multiplier and Likelihood Ratio Principles all lead to exactly the same test statistic in the case of a linear hypothesis, is re-examined in the light of a strict interpretation of these principles. Simple relations between various test statistics and their distributions are outlined. The concept of an artificial model is analyzed. A distinction is made between an artificial model that is in some sense an 'unrestricted' specification and one that is simply an algorithm. For non-nested hypotheses, an artificial model with prior information on the parameters is regarded as conforming to the Wald Principle. When arbitrary numerical methods are used as 'identifying' restrictions, the artificial model reduces to an algorithm since it cannot reasonably be 'accepted'.

PRINCIPLES AND METHODS IN THE TESTING OF ALTERNATIVE MODELS

- Gordon Fisher and Michael McAleer -

1. INTRODUCTION

There has been during the last decade a growing tendency in econometrics to pay increasing attention to problems of inference, as distinct from problems of estimation. This tendency is a natural outcome of more than twenty years of research and experience on the methods of estimation appropriate to the exigencies of economic data and econometric models. Yet the development of these methods was, to a large degree, specialized or even *ad hoc*, and it took a conscious and sustained effort to understand their essential unity under the twin principles of maximum likelihood and least squares. There is a danger that the same sort of tendency, namely to proliferate methods without regard for their underlying unity, will creep into the development of methods of testing. Indeed, we already have available a battery of different tests and algorithms to apply, or adapt, to particular problems -- with specialized names applying to each. In this light, it is well to recognize that there are indeed only a few principles available upon which testing may be based, and thereby to seek to emphasize the essential unity between tests based on the same principle and the relations that obtain between tests based on different principles. Recently, Breusch and Pagan (1980) have taken a step in this direction, using as their basis the Lagrange Multiplier Principle. The purpose this paper is to go a step further, by focusing on the various principles available and how these may apply to different problems that arise.

The paper is organized as follows. In Section 2 we distinguish between the principles and methods that may be used for testing economic hypotheses and comments are made upon them. The purpose is to develop a general framework for

discussion in the rest of the paper. Section 3 considers several examples in the context of nested hypotheses to highlight the usefulness of concentrating on a few principles of testing rather than on the numerous methods that are available. A test for specification error is considered in Section 4 as a means of introducing artificial regression models. The concept of an artificial model leads on naturally to testing non-nested hypotheses, so that Section 5 is devoted to the two principles and infinite number of methods that may be used. There follow, in Section 6, some concluding remarks.

2. PRINCIPLES AND METHODS

A distinction is to be drawn between the principles on which hypothesis testing is based and the practical statistical methods which these principles generate. The term *principle* denotes a general rule which specifies how tests are to be devised, while *method* signifies a specific statistical procedure arising from application of a principle to a particular problem. The distinction is helpful because it is common for many methods of testing to be devised on the basis of a single principle, but not *vice-versa*. The development of theory is then more straightforward and concise in terms of principles than in terms of methods, since the former avoids unhelpful repetition of notation and ideas. Moreover, knowledge that different methods have a common root in a particular principle is a useful aid to memory.

The usefulness of this distinction may be illustrated by reference to three common forms of testing nested hypotheses in large samples. These are Wald's (1943) test, Rao's (1948) test based on efficient scores, and the Lagrange-multiplier test (Aitchison and Silvey, 1958; Silvey, 1959), each of which was originally developed on the basis of maximum-likelihood theory. The second of these is exactly the same as the third, by virtue of first-order conditions on

the Lagrangean, and so the two will be considered as one. The main outcome of
the theory of these tests is that they all yield large-sample equivalents of
the likelihood-ratio test and corresponding estimators whose distributions are
almost always asymptotic normal. In consequence, any estimators that correspond
to these (*i.e.* which have distributions that are also asymptotic normal) may be
used to form corresponding tests. Similarly, since many standard tests arise as
a consequence of exact or approximate normality of the estimators involved, it is
to be expected that a whole range of standard methods are either straightforward
applications, or small-sample refinements, of the same tests.

Consider, for example, the estimation of a vector-valued parameter θ from
a random sample of n observations from a given distribution; θ is unknown, save
that it lies in p-dimensional Euclidean space Ω (p<n). It is desired to test
H_0: $\theta \in \omega$, where ω represents a sub-set of points in Ω which obey the r ($\leq$p)
restrictions $h(\theta) = 0$. If θ_Ω denotes maximum-likelihood estimate of θ in Ω (*i.e.*
unrestricted maximum-likelihood estimation), then the Wald (W-) test for H_0 is
given by

$$(2.1) \quad W = h^T(\theta_\Omega)[D_\Omega\{h(\theta_\Omega)\}]^{-1} h(\theta_\Omega) \quad ,$$

a standardized quadratic form in $h(\theta_\Omega)$, where $D_\Omega\{\cdot\}$ denotes dispersion matrix
corresponding to (unrestricted maximum-likelihood) estimation in Ω. Subject to
the usual regularity conditions, θ_Ω and $h(\theta_\Omega)$ will be asymptotic normal under H_0,
whence $W \overset{a}{\sim} \chi^2(r)$.[1] Similarly, if θ_Ω now refers to another asymptotic normal

---

[1]This statement will suffice for practical purposes. To be precise, $h(\theta_\Omega)$ is
consistent for $h(\theta)$ and it therefore has a degenerate distribution. Accordingly,
we must consider $\sqrt{n}\ (h(\theta_\Omega) - h(\theta)) \overset{d}{\to} N(0,\ HB^{-1}\ H^T)$, where $H = \partial h(\theta)/\partial\theta^T$, B is
the information matrix corresponding to 1 observation , and both H and B^{-1} are
evaluated at the true value of θ_Ω.

estimator of θ in Ω and $D_\Omega\{\cdot\}$ denotes its dispersion matrix, then W is again

a $\chi^2(r)$ variate under H_0; or if $h(\cdot)$ is linear and θ_Ω is unbiased and exactly

normal, then a small-sample refinement of W based on the F-distribution may be

obtained. We will return to this below.

Notice that, whatever W-test is used, its associated estimates are invar-

iably based upon unrestricted estimation, that is, upon estimation of θ in Ω,

disregarding the restrictions $h(\theta) = 0$. For this reason, we may associate the

Wald (W-) Principle with the notion of testing restrictions using standardized

quadratic forms of them based solely upon *unrestricted* estimation. In contrast,

the *Lagrange-multiplier (M-) Principle* is based solely upon estimation of θ in ω,

that is, upon *restricted* estimation, using f_ω, the estimate of the Lagrange mul-

tiplier, ϕ, corresponding to $h(\theta) = 0$. Of course, the large-sample test based

upon the M-principle is given by

$$(2.2) \qquad M = f_\omega^T \, [D_\omega(f_\omega)]^{-1} f_\omega \quad ,$$

where $D_\omega(\cdot)$ denotes dispersion matrix corresponding to (restricted) estimation

in ω, such that M is asymptotically distributed as $\chi^2(r)$ under H_0. Incidentally,

there is no need to insist on maximum-likelihood estimation: the estimated

Lagrange-multiplier f_ω may, for example, apply to least squares or some other

method of estimation, provided the estimates involved have well-defined normality

properties of the kind required.

Corresponding to the W- and M-principles we have the *Likelihood Ratio (L-)*

Principle which makes use of both restricted and unrestricted estimation. In

view of the bases of the tests, intuition would then suggest that application of

the W-principle will, in general, reject H_0 at least as often as application of

the M-principle, while application of the L-principle will lead to results that

lie somewhere in between the two. This is because unrestricted estimation corresponds to the case when H_0 is rejected, while restricted estimation corresponds to its 'acceptance'. In a sense, the use of both restricted and unrestricted estimators might be considered as an attempt to strike a 'balance' between the one and the other.

3. SOME APPLICATIONS: NESTED HYPOTHESES

We shall now consider particular applications of the principles introduced in Section 2 for the case of nested hypotheses.

Example 3.1: The Linear Hypothesis. Seber (1964) has investigated the testing of linear hypotheses in small samples according to the W-, M-, and L-principles and has concluded that all "...lead to exactly the same test statistic" (p. 265). While this conclusion is correct, Seber's method of establishing it does not conform to a strict application of the principles involved. The purpose of the following argument is to re-establish Seber's result while remaining faithful to the principles to be applied.

Consider the vector y which ranges over n-dimensional Euclidean space $\mathcal{E}_n$ according as $N(\mu, I_n \sigma^2)$. It is given that $\mu \in \Omega$, a p-dimensional sub-space, but otherwise μ and σ^2 are unknown. Corresponding to the sub-space Ω, the least squares estimates of μ and σ^2 are denoted by m_Ω and s_Ω^2, respectively. It is desired to test the linear hypothesis H_0: $\mu \in \omega$; $\omega \subset \Omega$, where ω is (p-r) dimensional. The number r represents the number of linear restrictions on Ω to define ω. Least squares estimation under H_0 yields m_ω and s_ω^2.

The standard test-statistic for H_0 is:

$$(3.1.1) \quad F = \frac{y^T(P_\Omega - P_\omega)y}{y^T(I_n - P_\Omega)y} \cdot \frac{n-p}{r}$$

where P denotes an orthogonal projection: P_Ω is on Ω along $\Omega^\perp$ and P_ω is on ω along $\omega^\perp$, orthogonal complementation $(^\perp)$ being relative to $\mathcal{E}_n$. More explicitly, if ω is defined by $\omega \equiv \Omega \cap N[A^T]$, where A^T is a known $r \times n$ matrix of rank $r \leqq p$, then any $x \in \Omega$ which obeys $A^T x = 0$ must lie in ω. Hence another statement of H_0 is: $A^T \mu = 0$, $\mu \in \Omega$. Corresponding to this latter statement, it is well known that the unique orthogonal projection on $\omega^\perp \cap \Omega$, namely $P_\Omega - P_\omega$, may be written as $P_\Omega A (A^T P_\Omega A)^{-1} A^T P_\Omega$, provided $R[A] \cap \Omega^\perp$ comprises the origin only (Seber, 1964, p. 262). It is then easy to demonstrate that F in (3.1.1) embodies the W-principle since $m_\Omega = P_\Omega y$ and hence

$$(3.1.2) \quad F = \frac{y^T (P_\Omega - P_\omega) y}{r s_\Omega^2} = \frac{(A^T m_\Omega)^T [D_\Omega (A^T m_\Omega)]^{-1} (A^T m_\Omega)}{r}$$

where $D_\Omega(\cdot)$ denotes dispersion matrix evaluated at $\sigma^2 = s_\Omega^2$, the latter being given by $s_\Omega^2 = \{1/(n-p)\} \{y^T (I_n - P_\Omega) y\}$. Of course, F in (3.1.1) and (3.1.2) each have the central $F(r, n-p)$ distribution under H_0. Further, rF is a quadratic form based upon the unrestricted estimates m_Ω and s_Ω^2 and the given restrictions only; upon replacing s_Ω^2 with σ^2, it is seen to be a quadratic form in standard-ized normal variates under H_0. Since also s_Ω^2 is asymptotically equivalent to σ_Ω^2, the maximum-likelihood estimator of σ^2, it is obvious that $rF \sim \chi^2(r)$ for large n. Indeed, if s_Ω^2 is replaced by σ_Ω^2, we may write $rF = W$ to comply with the original definition of W in (2.1).

The corresponding small-sample test for H_0 based upon the M-principle may be obtained *via* minimization of $(y-\mu)^T (y-\mu)$ subject to $A^T \mu = 0$ for $\mu \in \Omega$. This requires finding a stationary point on

$$(3.1.3) \quad L = (y-\mu)^T (y-\mu) + 2\mu^T A \phi - 2\mu^T (I - P_\Omega) \kappa$$

for variations in μ and the vector Lagrange multipliers ϕ and κ which minimizes

$(y-\mu)^T(y-\mu)$ while satisfying $A^T\mu=0$ for some $\mu \in \Omega$. The small-sample application of the M-principle is based upon the estimate of ϕ from (3.1.3) and the implicit hypothesis corresponding to H_0, namely: $\phi = 0$. Note carefully that the entire procedure is based upon least squares estimation of μ in ω. Writing f_ω for the estimate of ϕ corresponding to m_ω, the first-order conditions from (3.1.3) lead to the small-sample test statistic based upon the M-principle:

$$(3.1.4) \quad M = f_\omega^T \left[D_\omega(f_\omega) \right]^{-1} f_\omega = y^T P_\Omega A (A^T P_\Omega A)^{-1} \left[(A^T P_\Omega A)^{-1} s_\omega^2 \right]^{-1} (A^T P_\Omega A)^{-1} A^T P_\Omega y \quad ,$$

where $s_\omega^2 = y^T(I_n-P_\omega)y/(n-p+r) = (y-m_\omega)^T(y-m_\omega)/(n-p+r)$.

Seber's (1964) demonstration that both the W- and the M-principles lead to exactly the same test statistic lies in noting that, if σ_Ω^2 replaces s_ω^2 in the second term in (3.1.4), the resulting expression is $\{nr/(n-p)\}$ times the F-statistic of (3.1.1). Unfortunately, this step involves evaluating the dispersion matrix of f_ω, namely $(A^T P_\Omega A)^{-1} \sigma^2$, at the *unrestricted* estimate σ_Ω^2. It is clear this violates the M-principle as established above, since this principle requires the use of restricted estimates only. Moreover, using σ_Ω^2 in place of s_ω^2 is unnecessary to demonstrate that the two principles lead to the same F-statistic, as the following argument reveals.

Equation (3.1.4) is readily seen to reduce to

$$(3.1.5) \quad M = \frac{y^T(P_\Omega-P_\omega)y}{y^T(I_n-P_\omega)y} \cdot (n-p+r) \quad .$$

Under H_0, $\{M/(n-p+r)\}$ is distributed as $\beta_1(\frac{r}{2}, \frac{n-p}{2})$ exactly, since

$$(3.1.6) \quad \frac{y^T(P_\Omega-P_\omega)y}{y^T(I_n-P_\omega)y} = \frac{y^T(P_\Omega-P_\omega)y}{y^T(P_\Omega-P_\omega)y + y^T(I_n-P_\Omega)y}$$

and the two components in the denominator of the right-hand side of (3.1.6), each divided by σ^2, are independent chi-square variates with r and (n-p) degrees of freedom, respectively.

There is, of course, a direct correspondence between the $\beta_1(\frac{r}{2}, \frac{n-p}{2})$ distribution and the central $F(r, n-p)$ distribution. If, for example, $v \sim \beta_1(\frac{q}{2}, \frac{m}{2})$ and $v = u/(1 + u)$, then $u = v/(1-v)$ and $u \sim \beta_2(\frac{q}{2}, \frac{m}{2})$; moreover, mu/q has the central $F(q,m)$ distribution. Thus although F and M will yield different calculated numbers in a practical example, there will be no conflict in using them to test H_0 since they have different, though corresponding, distributions. The relation between $W = rF$, of equations (3.1.1) and (3.1.2), and M, in equation (3.1.5), may be written

$$(3.1.7) \quad \frac{W}{n-p + W} = \frac{M}{n-p+r}$$

(see *e.g.* Weatherburn, 1952, chapter 8; Wilks, 1962, p. 187). Moreover, if λ is the likelihood ratio corresponding to H_0, it must depend on the values of the likelihoods corresponding to estimation in Ω and ω. Thus λ is based upon information contained in *both* W and M. This is readily seen from the definition of λ: $\lambda^{2/n} = \{\sigma_\Omega^2/\sigma_\omega^2\}$ where σ_ω^2 refers to maximum-likelihood estimate of σ^2 in ω. Thus, for large n, the relations

$$\lambda^{2/n} = \{M/W\} = \{s_\Omega^2/s_\omega^2\}$$

hold approximately, whereas for any finite n, the following holds exactly:

$$(3.1.8) \quad \lambda^{2/n} = \frac{M}{W} \cdot \frac{n-p}{n-p+r} \quad .$$

Note also that

$$(3.1.9) \quad W = (\lambda^{-2/n} -1)(n-p) \quad .$$

Hence, there can be no conflict between the small-sample tests based upon the W- and L-principles. It follows immediately that there can be no conflict between the small-sample refinements of tests based upon the W-, M- and L-principles.

With regard to the calculated values of the test statistics, it is clear from (3.1.6), (3.1.7) and (3.1.9) that

$$(3.1.10) \quad W = (\lambda^{-2/n} - 1)(n-p) \geqq M\{(n-p)/(n-p+r)\}$$

which may be regarded as the exact small-sample relation between the three tests corresponding to the general large-sample relation:

$$(3.1.11) \quad W\{\frac{n}{n-p}\} \geqq \{-2 \log \lambda\} \geqq M\{\frac{n}{n-p+r}\} \quad ,$$

each of which has the $\chi^2(r)$-distribution for large n. Relations (3.1.7) - (3.1.9) admit of proper application of the principles involved and we see that, while the calculated values of the W- and M- statistics will differ, there is no conflict between the tests since each is based upon its own distribution. Finally, since there is a one-for-one correspondence between λ and W, all three principles are seen to lead to the same test statistic; for convenience, this may be taken as the F- statistic given in (3.1.1).

Finally, notice that, in the linear case examined here, the essential difference between a test based upon the W-principle and one based upon the M-principle lies in the estimate of σ^2 that should be used. In the former case, an estimate of σ^2 in Ω is required, namely σ_Ω^2 or s_Ω^2 in the above example; when the M-principle is applied, it is necessary to use an estimate of σ^2 in ω.

Example 3.2: Classical Identifiability. Consider an equation, say the first, of a standard interdependent linear system in n observations on M endogenous

variables y_i, K exogenous variables x_j and M jointly normal equation errors ε_i, which may be written:

$$(3.2.1) \quad y_1 - Y_1\beta_1 = X_1\gamma_1 + \varepsilon_1 \quad .$$

This equation comprises $(m + 1)$ endogenous variables, $[y_1:y_2\cdots y_{m+1}]=[y_1:Y_1]=Y_1^*$ and k exogenous variables $[x_1 x_2 \cdots x_k] = X_1$. From the reduced form of the system, writing $\beta_1^{*T} = [1:-\beta_1^T]$, we have

$$y_1^* = y_1 - Y_1\beta_1 = Y_1^*\beta_1^* = X_1\Pi_{11}\beta_1^* + X_2\Pi_{21}\beta_1^* + V_1\beta_1^* \quad ,$$

the Π's being appropriate matrices of reduced form coefficients, V_1 the matrix of reduced form errors corresponding to Y_1^*, and $X_2 = [x_{k+1}, x_{k+2}, \ldots, x_K]$. It is presumed that $K-k > m$, so the first equation is over-identifiable. The over-identifying restrictions imply $\Pi_{11}\beta_1^* = \gamma_1$, $\Pi_{21}\beta_1^* = 0$ and $V_1\beta_1^* = \varepsilon_1$. If β_1^* is presumed known, the over-identifying restrictions may be tested, using the results in *Example 3.1*. Noting that Ω is the span of $X = [X_1:X_2]$ while ω is the span of X_1, we have

$$W = \frac{y_1^{*T}(P_\Omega-P_\omega)y_1^*}{y_1^{*T}(I_n-P_\Omega)y_1^*} \cdot \frac{n-K}{K-k} \quad ,$$

which has the central $F(K-k, n-K)$ distribution under the null hypothesis $\Pi_{21}\beta_1^* = 0$. The corresponding M-test is

$$M = \frac{y_1^{*T}(P_\Omega-P_\omega)y_1^*}{y_1^{*T}(I_n-P_\omega)y_1^*} \cdot (n-k) \quad ,$$

where $\{M/(n-k)\}$ has the beta type 1 distribution with parameters $(K-k)/2$ and $(n-K)/2$ under the same null hypothesis (see Anderson and Rubin, 1949, 1950;

Fisher, 1977). Unfortunately, the parameter β_1 is not known. If β_1 is estimated by limited-information maximum-likelihood (LIML), then β_1^* is selected by minimizing W with respect to β_1^* for given n, k and K, whence it is seen to be independent of the normalization on β_1^*. In these circumstances there are m additional restrictions placed on the data and Basmann (1960) recommends adjusting W by $\{(K-k)/(K-k-m)\}$ and using $F(K-k-m, n-K)$ as the approximate distribution under the null hypothesis. The approximation is evidently very good (Basmann, 1960; and *e.g.* Fisher, 1980). Notice that while a degrees-of-freedom adjustment is made for the factor $y_1^{*T}(P_\Omega - P_\omega)y_1^*$ of W, none is made for $y_1^{*T}(I_n - P_\Omega)y_1^*$, when estimates replace β_1. This is because the second factor divided by n, or $(n-K)$, or $(n-K-m)$ will be consistent for σ_{11}, the variance of ε_1; for large n, each divisor is dominated by n and hence finite adjustment is of no consequence. On the contrary, the matrix $(P_\Omega - P_\omega)$, of the first factor, has finite rank for all n and hence finite adjustment is of greater concern. By a similar argument, we might regard $\{M/(n-k)\}$ as an approximate beta type 1 distribution under the null hypothesis with parameters $(K-k-m)/2$ and $(n-K)/2$ when LIML estimates replace β_1 in the formula. Clearly, there is again a correspondence between the M- and the W- statistics, as in *Example 3.1*, except that now approximations are involved.

Example 3.3: Identifiability Tests with Two-Stage Least Squares (2SLS). If β_1 is estimated by 2SLS, it is selected to minimize the numerator of M or W, in which case Fisher (1980) has shown that W reduces to

$$\{e_1^T P_\Omega e_1/(K-k-m)\}/\{e_1^T(I_n - P_\Omega)e_1/(n-K)\} \quad ,$$

where e_1 is the 2SLS residual from (3.2.1). This result arises because $P_\omega e_1 = 0$. Hence, M may be seen to reduce to $\{e_1^T P_\Omega e_1\}/\{e_1^T e_1/(n-k)\}$. These are extremely simple statistics to calculate since they merely involve the artificial regres-

sion of e_1 on the columns of X to obtain $e_1^T (I_n - P_\Omega) e_1$, from which $e_1^T P_\Omega e_1$ may also be obtained. Moreover, corresponding statistics may be developed to check 'identifiability' when applying instrumental variable estimation. Note that no maximum-likelihood estimation is involved here. We merely make use of the asymptotic normality properties of 2SLS estimates. Although the $F(K-k-m, n-K)$ approximate distribution for W is close, nothing is known about the adequacy of the beta distribution approximations associated with M; however, there is a natural correspondence between the two, as before.

Example 3.4: 2SLS Significance Tests. If the null hypothesis is 'accepted' in the previous example, it may still be the case that further zero identification restrictions are in order; or it may be desirable to test whether certain parameter estimates deviate significantly from the values theory prescribes for them. In the case of significance tests of this kind, two routes are open: either the F-tests proposed by Dhrymes (1968), or their χ^2-(or asymptotic normal) equivalents. Morgan and Vandaele (1974) produce evidence in favour of the latter. Be this as it may, the only difference between the F- and the χ^2- statistics lies in the estimate of σ_{11} that is used: the former uses $\{e_1^T P_\Omega e_1 / (K-k-m)\}$ whereas the latter uses $e_1^T e_1 / n$. Both statistics are based upon the unrestricted regression (3.2.1) since, having established (over-) identifiability, the test is designed to consider further restrictions on the parameters β_1 and/or γ_1. Thus, both are based upon the W-principle. Note, however, that the estimate of σ_{11} arising from the M-test of *Example 3.3* is $e_1^T e_1 / (n-k)$, or $e_1^T e_1 / (n-k-m)$, both of which are asymptotically equivalent to $e_1^T e_1 / n$. Thus, the distinction between the F- and χ^2-statistics is not one of principle, but rather one of practice: which estimate of σ_{11} to use? Given this choice, there is also the implication regarding approximate distribution. Which choice to make is essentially an empirical issue. If $\hat{\sigma}_{11}$

denotes estimate of σ_{11}, whichever is used, and we consider only restrictions on γ_1, then the formulae for the tests is based upon the ratio

$$\frac{y_1^{*T}(P_\omega - P_r)y_1^*}{\hat{\sigma}_{11}} = \frac{y_1^{*T}(I_n - P_r)y_1^* - y_1^{*T}(I_n - P_\omega)y_1^*}{\hat{\sigma}_{11}} \quad ,$$

where P_r represents projection on the restricted regression space, which must be a sub-space of ω. If the corresponding restricted sum of squared residuals is $e_r^T e_r$, the last ratio becomes $(e_r^T e_r - e_1^T e_1)/\hat{\sigma}_{11}$, so that, if the M-principle is applied, $\hat{\sigma}_{11}$ must be based upon $e_r^T e_r$ rather than on $e_1^T e_1$.

Example 3.5: *Identifiability and the Reduced Form.* The unrestricted reduced form of a standard linear system may be estimated efficiently by single-equation least squares. Suppose the coefficients are represented by the vector π, then the over-identifying restrictions imply generally non-linear relations between the coefficients of different equations. These may be written $h(\pi) = 0$, $h(\pi)$ being vector-valued with as many components as there are 'extra' restrictions. If $\hat{\pi}$ represents the unrestricted least squares estimates of π, then an appropriate identifiability test statistic is

$$W = h^T(\hat{\pi})[H(\hat{\pi})\{\hat{\Sigma} \otimes (X^TX)^{-1}\} H^T(\hat{\pi})]^{-1}h(\hat{\pi}) \quad ,$$

where $H(\hat{\pi})$ is $[\partial h(\pi)/\partial \pi^T]$ evaluated at $\pi = \hat{\pi}$ and $\hat{\Sigma} = [\hat{\sigma}_{ij}]$, $\hat{\sigma}_{ij}$ being the mean of cross-products of residuals from equations i and j. This is a straightforward example of linear estimates being used to test non-linear constraints. W is a Wald statistic in the usual large-sample sense, and hence is distributed as $\chi^2(r)$, where r is the total number of restrictions in excess of the minimum required for identifiability.

Example 3.6: Structural Change. Consider a regression which is subject to structural change. Various tests are available, including the analysis of co-variance for the non-singular case, the prediction interval test for the corresponding singular case (see *e.g.* Chow, 1960), and the cusum of squares test using recursive residuals (Brown *et al*, 1975). The purpose here is to consider the Chow and cusum of squares tests, and the relations between them, in the light of the discussion in Sections 2 and 3.1.

If y_t is the dependent variable at time t and ε_t is the zero mean regression error at t, each being independently distributed normal variates over time with common variance σ^2, then

$$y_t = x_t^T \beta_t + \varepsilon_t \quad , \quad t-1,2,\ldots,p,p+1,\ldots,n$$

there being p components in the vectors x_t and β_t, the latter being unknown. It is presumed that the x_t are fixed in repeated samples. Let X_q denote the matrix of rows x_t^T, t=1,2,...,q, and consider testing $\beta_{p+1} = \beta_{p+2} = \ldots = \beta_n$. For this, we shall need orthogonal projections on the span of

$$\hat{X}_q = \begin{bmatrix} X_q & 0 \\ 0 & I_{n-q} \end{bmatrix} \quad , \quad q = p+1,\ p+2,\ \ldots,\ n$$

of rank (p+n-q), which is denoted by $\hat{P}_q$. The prediction interval test is based on the W-principle and, for $p+1 \leqq q \leqq n$, uses the statistic

$$W = \frac{y^T(\hat{P}_q - \hat{P}_n)y}{y^T(I_n - \hat{P}_q)y} \cdot \frac{q-p}{n-q} \quad .$$

W represents the Chow test, which may be repeated until a significant sum of squares is obtained for q = p+1, p+2, ..., n. Under the null hypothesis, W has the central F(n-q, q-p) distribution. Now consider the cusum of squares test

$$Q = \frac{y^T(I_n - \hat{P}_q)y}{y^T(I_n - \hat{P}_n)y} \quad ,$$

which is very simply related to a corresponding test statistic based on the M-principle (*c.f.* Fisher, 1978):

$$1 - Q = \frac{y^T(\hat{P}_q - \hat{P}_n)y}{y^T(I_n - \hat{P}_n)y} = \frac{M}{n-p} \quad .$$

Thus, the Chow test statistic is based on the W-principle and the cusum of squares statistic on the M-principle. Of course, the two correspond, which perhaps explains the assertion in Breusch and Pagan (1980, p. 251) that the Chow test is an M-statistic. The relation between the Chow and the cusum of squares statistics is given by

$$WQ \cdot \frac{n-q}{q-p} = 1 - Q \quad .$$

Since there can be no conflict in the outcomes of these tests, choice between them is a matter of convenience. However, in applications of the TIMVAR programme (Brown *et al*, 1975), it should be noted that the exact distribution of Q is not used. Hence, there may be conflict between the results emerging from TIMVAR calculations and corresponding calculations using the exact distribution of Q or W.

4. *CONSTRUCTION OF AN ARTIFICIAL MODEL*

The principles outlined in Section 2 may be applied to other problems that arise in econometrics. In this section we outline a large-sample test of mis-specification as a means of introducing artificial regression equations.

Consider the standard linear regression equation, $y = X\beta + \varepsilon$. An important assumption is that $E(\varepsilon/X) = 0$ or, in large samples, that $\text{plim } \{(1/n)X^T\varepsilon\}=0$. Situations in which the assumption may not reasonably be maintained are when X is subject to measurement error, when X contains a lagged dependent variable in the presence of serially correlated errors, or when X contains endogenous variables.

We may formulate the problem of mis-specification by setting $\beta = \beta^* + \delta$, where β^* is the true value of β and δ is the specification bias. Under the null hypotheses $\delta=0$, whereas $\delta \neq 0$ under the alternative. Denote by $\hat{\beta}_0$ and $\hat{\beta}_1$ the OLS and instrumental variable (IV) estimators of β, respectively. Then $\hat{\beta}_0$ is consistent and asymptotically efficient under H_0 but inconsistent under H_1, while $\hat{\beta}_1$ is consistent under both hypotheses. A test of mis-specification consists of examining how large $\hat{\delta} = \hat{\beta}_1 - \hat{\beta}_0$ is in relation to its variance, $V(\hat{\delta})=V(\hat{\beta}_1)-V(\hat{\beta}_0)$. Under the null hypothesis, the statistic

$$(4.1) \quad (\sqrt{n}\hat{\delta})^T \{\hat{V} (\sqrt{n} \ \hat{\delta})\}^{-1} (\sqrt{n} \ \hat{\delta}) \overset{a}{\underset{\sim}{}} \chi^2(r) \quad ,$$

where r is the number of unknown elements of β (see Durbin, 1954). In practice, $V(\hat{\delta})$ is consistently estimated by $\hat{V}(\hat{\delta}) = \hat{\sigma}_1^2(X^TP_zX)^{-1} - \hat{\sigma}_0^2(X^TX)^{-1}$, where P_z is the orthogonal projection on the span of Z, the matrix of observations on the set of instrumental variables, and $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ are the OLS and IV estimates of the error variance. Under the null hypothesis, $\hat{\sigma}_0^2$ may be substituted for $\hat{\sigma}_1^2$, and *vice-versa*, since any consistent estimate of σ^2 may be used to calculate (4.1).

An alternative test of mis-specification arises by considering the components of $X\beta$ in the span of Z and its orthogonal complement, and recognizing that OLS will yield consistent estimates of β^* for the former and β for the latter, under the alternative hypothesis, and β^* for both under the null. This

way of looking at the problem is due to Hausman (1978) and leads to consideration of the artificial regression equation

$$(4.2) \quad y = X\beta + P_z X\delta + \upsilon \quad .$$

Under the null hypothesis $\delta=0$, the test from (4.2) is asymptotically equivalent to the test given by (4.1). Since the columns of X must first be regressed on the columns of Z to obtain $P_z X$, the test implicit in (4.2) is not necessarily easier to compute than is (4.1). The artificial model (4.2) is of interest as an *algorithm* for testing $\delta=0$; it is simply a method for calculating a test statistic which is asymptotically equivalent to (4.1). Although $y = X\beta + \varepsilon$ is consistent with $\delta=0$, the artificial specification given by (4.2) is not necessarily the only one consistent with $\delta\neq0$.

While artificial models constructed for the express purpose of calculating a test statistic may be regarded as conforming to the W-principle, in that they are indeed 'unrestricted' forms, the artifact itself is not the subject of interest. Nevertheless, the distinction between the W- and M-principles does enable various methods of testing to be developed. The concept of an artificial model is of particular interest because it is linked with various procedures available for testing non-nested hypotheses, to which we now turn.

5. *NON-NESTED HYPOTHESES*

We may think of two hypotheses as being non-nested when the specification of one cannot be obtained from the other by the imposition of appropriate restrictions, or be obtained as a limiting form of a suitable approximation. The testing of non-nested hypotheses has become increasingly popular following the adaptation of the ideas of Cox (1961, 1962) to econometrics by Pesaran (1974) and

Pesaran and Deaton (1978). For an interpretation of these tests, particularly whether the tests should be one- or two-sided, see Fisher and McAleer (1980a). While (1980) provides the general regularity conditions which ensure that the Cox-statistic is asymptotic normal with zero mean and unit variance. Davidson and MacKinnon (1980) have developed several algorithms for testing non-nested hypotheses and have related these to the Cox-test. Fisher and McAleer (1980b) derived variations of both the Cox-test and Davidson and MacKinnon's J-test, using an idea implicit in Cox (1961) and exploited by Atkinson (1970), and demonstrated that the Atkinson variation always provides a consistent test (*c.f.* Pereira, 1977). In this section we examine two principles for *testing* non-nested hypotheses, namely the Modified Likelihood Ratio (MLR) Principle of Cox and the Principle of Artificial Nesting (AN) (see *e.g.* Hoel, 1947), and the various methods that may be used in each case. We also briefly outline a method of *choosing* between non-nested regression models which will lead to the selection of the correct model, 'on average', as long as one of the models being considered is 'true' (see *e.g.* Theil, 1961).

We shall restrict ourselves to the class of non-linear, unilateral causal dependancies and follow closely, but not completely, the notation of Pesaran and Deaton (1978). For simplicity we consider only single-equation models; extensions to systems of non-linear equations of the same type are straightforward. Suppose there are two competing economic theories to explain the behaviour of an endogenous variable y, and these are written:

$$(5.1) \quad H_0: \quad y = f(\theta_0; X) + u_0, \quad u_0 \sim N(0, I_n \sigma_0^2)$$

and

$$(5.2) \quad H_1: \quad y = g(\theta_1; Z) + u_1, \quad u_1 \sim N(0, I_n \sigma_1^2)$$

where y is an n-vector of co-ordinates y_i, and $f(\theta_0;X)$ and $g(\theta_1;Z)$ denote n-vectors of co-ordinates $f(\theta_0;x_i)$ and $g(\theta_1;z_i)$, respectively, x_i and z_i being the i'th rows of matrices of exogenous observations X and Z. The functions $f(\cdot)$ and $g(\cdot)$ are continuous and at least twice differentiable with respect to all their arguments. It is assumed that the vector-valued parameters θ_0 and θ_1 are identifiable under H_0 and H_1, respectively. We shall write $f(\theta_0;X)$ variously, according to convenience, as $f(\theta_0)$ or simply f; similarly for $g(\theta_1;Z)$.

Let us write α_0^T and α_1^T for the parameter sets (θ_0^T, σ_0^2) and (θ_1^T, σ_1^2), respectively. The likelihood functions under H_0 and H_1 are:

$$(5.3) \quad L_0 \equiv L_0(\alpha_0) = (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_0^2}(y-f)^T(y-f)\right\}$$

and

$$(5.4) \quad L_1 \equiv L_1(\alpha_1) = (2\pi\sigma_1^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_1^2}(y-g)^T(y-g)\right\}.$$

Given the nature of the distributions under H_0 and H_1, it is natural to consider an exponential weighting scheme (*c.f.* Quandt, 1974, where a linear combination of the likelihood functions is analyzed). The exponential combination of L_0 and L_1, involving the nesting parameter λ, is given by:

$$(5.5) \quad L \equiv L(L_0,L_1;\lambda) = \kappa\, L_0^{1-\lambda} L_1^{\lambda}$$

where κ, the factor of proportionality, is given by $\left\{\int_{-\infty}^{+\infty} L_0^{1-\lambda} L_1^{\lambda}\, dy\right\}^{-1}$. The weighted likelihood function given by (5.5) is important for two reasons.

(*i*) The first-order maximizing condition with respect to λ leads to the numerator of the Cox-statistic (Atkinson, 1970, p. 334; see also Breusch and Pagan, 1980, p. 248). Equivalently, a Lagrange multiplier may be introduced explicitly. In this sense, the Cox test is seen to be based upon the M-principle. Hence this and the MLR-principle are seen to yield the same test

in this case. However, the nesting parameter λ is not estimated under the MLR-principle, since the Cox statistic is evaluated under H_0 or H_1, whereupon λ is set *a priori* to zero or unity.

(*ii*) According to the AN-principle, on the contrary, λ may itself be elevated to the position of test statistic and its estimation must be attempted. As the introduction of λ involves an 'unrestricted' artificial model, tests based on the AN-principle would also seem to be based on the W-principle. Unfortunately, λ is not identifiable, in general, and may be made identifiable only under specialized circumstances. Since there are numerous ways of imposing 'identifying' restrictions, the principle on which a particular method is based will vary with the information actually used. We return to this below.

5.1 *The Modified Likelihood Ratio*

The Cox-test under H_0 is given by $N_0 = T_0/\{\hat{V}_0(T_0)\}^{\frac{1}{2}}$, where N_0 is asymptotically distributed as $N(0,1)$ under H_0 (see Cox, 1961, 1962; for the general regularity conditions, see White, 1980). The first-order maximizing condition with respect to λ in the weighted likelihood function (5.5) leads to the numerator of the Cox-statistic (Atkinson, 1970, p. 334; see also Breusch and Pagan, 1980, p. 248); hence, T_0 is given by:

$$(5.1.1) \quad T_0 = \frac{n}{2} \log (\hat{\sigma}_1^2/\hat{\sigma}_{10}^2)$$

where $\hat{\sigma}_{10}^2 = \hat{\sigma}_0^2 + \hat{\sigma}_a^2$, $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ are the maximum likelihood estimates of σ_0^2 and σ_1^2, respectively, and $\hat{\sigma}_a^2 = \{(\hat{f}-g(\hat{\theta}_{10}))^T(\hat{f}-g(\hat{\theta}_{10}))/n\}$ is the estimated error variance from the auxiliary regression of $\hat{f} \equiv f(\hat{\theta}_0)$ on the model under H_1 (see Pesaran and Deaton, 1978, for details). In calculating T_0, $\hat{\theta}_1$ has been used for θ_1. This is an arbitrary substitution since, when H_0 is under test, the maximized value of L_0 is independent of the value assigned to θ_1. Therefore, under H_0, Atkinson (1970) advocates replacing θ_1 with $\hat{\theta}_{10}$, a consistent estimate

of the asymptotic expectation of $\hat{\theta}_1$ under H_0, rather than with $\hat{\theta}_1$, the estimate of θ_1 under H_1. This is an attractive way of looking at the problem, not only because it entails evaluating the entire statistic under H_0, but also because it leans more heavily in the direction of H_0 than if $\hat{\theta}_1$ is used. It should be noted, however, that the latter statement holds only when H_0 is performing better than expected. Should H_0 then be rejected by H_1, this would seem to provide stronger evidence against H_0 than if $\hat{\theta}_1$ had been used. Fisher and McAleer (1980b) have demonstrated that this is indeed the case, and have derived two variations of the Cox-statistic that are asymptotically equivalent to T_0 under H_0. The relation between them is $TA_0 \geqq TL_0 \geqq T_0$, where TA_0 and TL_0 are given by

$$(5.1.2) \quad TL_0 = \frac{n}{2}\{(\hat{\sigma}_1{}^2/\hat{\sigma}_{10}{}^2) - 1\}$$

and

$$(5.1.3) \quad TA_0 = TL_0 + \frac{1}{2\hat{\sigma}_{10}{}^2}\{(y-g(\hat{\theta}_{10}))^T(y-g(\hat{\theta}_{10})) - e_1{}^T e_1\}.$$

As $e_1{}^T e_1$ is the minimized sum of squares under H_1, the second term in (5.1.3) will be positive unless the model under H_0 fits perfectly (*i.e.* $g(\hat{\theta}_{10}) = g(\hat{\theta}_1)$).

Although the choice of the asymptotic variance of TA_0 and TL_0 is arbitrary, it will be convenient to use the estimate of $V_0(T_0)$ given by Pesaran and Deaton (1978, p. 687). Dividing TA_0, TL_0 and T_0 by $\{\hat{V}_0(T_0)\}^{\frac{1}{2}}$ leads to $NA_0 \geqq NL_0 \geqq N_0$. It is obvious that when the alternative, H_1, is fitting much better (worse) than it ought, relying solely on $N_0(NA_0)$ will more likely lead to rejection of H_0 than would otherwise be the case. The linearized Cox-statistic, NL_0, is therefore more conservative with respect to rejecting the model under test than is $N_0(NA_0)$, when the alternative is fitting much better (worse) than might be expected. It may prove useful to compute all three test statistics since differences in their numerical values, even for moderately large samples, has served

to guide the interpretation underlying the rejection of the tested hypothesis (see Fisher and McAleer, 1980b).

The three statistics should converge in probability under H_1 to a negative value, since each should have a negative mean when H_1 is at work. For a test to be consistent, the probability of rejecting H_0, when any member of the class of alternatives H_1 holds, must tend to one as the sample size increases indefinitely. Pereira (1977) has shown that the numerator of the Cox-test will always converge in probability to a negative value under H_1, whereas the modification suggested by Atkinson (1970) *may* converge to a positive value. In this way, the Cox test is always consistent; but the Atkinson variation of it derived above may lead to an inconsistent test. This result serves to illustrate the fact that, while two statistics may be asymptotically equivalent under the null hypothesis, they need not be under the alternative. For the case of non-nested, non-linear regression models, Fisher and McAleer (1980b) have demonstrated that the Atkinson variation of the Cox test always provides a consistent test. However, since the variances of the Cox statistics are not equal under the alternative hypothesis, it does not seem possible to show which of the three statistics has greatest power.

It was stated above that the Cox test was based upon the M-principle. However, the very notion of modifying a likelihood ratio suggests that the MLR-principle is an adjunct to the Likelihood Ratio (L-) principle. The original Cox test of Pesaran and Deaton (1978) uses estimates from both H_0 and H_1, and is in accordance with the L-principle. On the other hand, the Atkinson variation evaluates the entire statistic under H_0 and, from this point of view, adopts the M-principle. As in the case of nested hypotheses, the two principles lead to test statistics that are asymptotically equivalent under the tested hypothesis.

5.2 *Artificial Nesting*

The following artificial regression model is implicit in (5.5):

$$(5.2.1) \quad y = \left\{ \frac{(1-\lambda)\sigma_1^2}{(1-\lambda)\sigma_1^2 + \lambda\sigma_0^2} \right\} f + \left\{ \frac{\lambda\sigma_0^2}{(1-\lambda)\sigma_1^2 + \lambda\sigma_0^2} \right\} g + u$$

where the variance of u is given by $\sigma^2 = \{\sigma_0^2 \sigma_1^2\}/\{(1-\lambda)\sigma_1^2 + \lambda\sigma_0^2\}$. Note that λ must be restricted to ensure that σ^2 is positive; this condition is certainly satisfied when $\lambda \in [0,1]$. The artificial model is a linear combination of the models under H_0 and H_1, with weights depending explicitly on the nesting parameter and the error variances σ_0^2 and σ_1^2. This has been examined for the linear case by Atkinson (1970) and Quandt (1974), while Pesaran (1979) and Fisher and McAleer (1980b) considered the non-linear case. It is important to realize that λ, the parameter of greatest interest in (5.2.1), is not, in general, identifiable. Seen in this light, H_0 and H_1 cannot be tested without the imposition of appropriate identifying restrictions. In the event that no such restrictions are forthcoming, we would simply be left with the 'unrestricted' model (5.2.1). Pesaran (1979) has considered making λ identifiable by assuming the parameters of H_1 are available. We will briefly outline below different 'identifying' restrictions that will enable H_0 and H_1 to be tested.

(i) *A Priori Identification*

If both α_0 and α_1, and hence f and g, are *known* in (5.2.1), we can test H_0 and H_1 by testing λ against zero and unity, respectively. But since (5.2.1) is non-linear in λ, numerical optimization would be necessary to estimate it. If we take the view that the hypotheses under test relate only to the expected value of y, we may quite reasonably impose $\sigma_0^2 = \sigma_1^2 = \sigma^2$. Equation (5.2.1) then simplifies to:

$$(5.2.2) \quad y-f = \lambda(g-f) + u \quad .$$

Since λ is identifiable in (5.2.2), it is both the nesting and the testing parameter. An alternative method of tackling the problem is to make a distinction between the nesting and the testing parameters. After appropriate reparametrization, (5.2.1) may be written as:

$$(5.2.3) \quad y-f = \Lambda(g-f) + u$$

where $\Lambda = \{\lambda\sigma_0^2\}/\{(1-\lambda)\sigma_1^2 + \lambda\sigma_0^2\}$. The testing parameter is now denoted by Λ, and this is to be distinguished from the nesting parameter λ. In order to test the two hypotheses in (5.2.3), prior information is required only for θ_0 and θ_1, since the variances are subsumed in Λ. In respect of (5.2.2), least squares estimation of λ yields a t-ratio with (n-1) degrees of freedom upon which the test may be based (see Hoel, 1947; Fisher and McAleer, 1980a).

Insofar as λ is non-zero, different values of λ have different interpretations with regard to movement toward, or away from, the alternative hypothesis. Thus, in respect of H_0, λ as a positive fraction is interpreted as a movement from H_0 toward H_1, and a value greater than unity is interpreted as a value 'beyond' H_1. On the contrary, a negative value of λ is a movement away from both H_0 and H_1. Corresponding interpretations may be given for H_1 when λ differs in value from unity. It is entirely possible that one model is better than another, although a linear combination of the two formulae may be better still. This arises when there are significant departures from each model toward the other, corresponding to $0 < \lambda < 1$. It must be recognized that an exponential weighting of the likelihood functions of two non-nested models inextricably creates an artificial model which may be 'acceptable' to the data when the sample is consistent with neither H_0 nor H_1. Application of the AN-principle

therefore has the distinguishing feature that the artificial model may be 'accepted', so that an artificial model with prior information on the parameters conforms to the W-principle. As for application of the MLR-principle, there is no artificial model as such, so the question of its 'acceptance' never arises.

A priori information on the parameters of the competing models will not generally be available for applying the Hoel-test. However, given the applicability of independent observations, a forecasting formula may be "...obtained from a combination of theoretical considerations and past observations..." (Hoel, 1947, p. 605). As a device for acquiring information, we may use 'prior' in the intertemporal sense by partitioning the n independently normally distributed random variables into two (not necessarily equal) regimes, using the observations from one regime to obtain two calculated forecasting formulae, and then testing the formulae thus obtained on the observations of the second regime. The process may be reversed, in which case the formulae obtained from the second regime may be tested on the first. If the predictive performance of a model or formula is of primary concern, we may simply construct the formula using the data from the first regime, and test the ability of the formula to provide good predictions with the data of the second regime. In this sense, the Hoel-test is analogous to the sequentially-repeated F-test for structural change (see, *e.g.*, Chow, 1960), and tests in the tradition of data analysis as put forward by, for example, Brown *et al* (1975) (see Section 3.6 above).

The use of sample information from one regime is extraneous with respect to the other regime and, although there is a degree of arbitrariness involved in the selection of the observations for the two regimes, this is of no particular

consequence if the observations are independent, as is presumed. Of course, in many cases in econometrics this presumption is unwarranted. However, having completed a limited number of Monte Carlo experiments, it is satisfying to find, in comparing a 'true' linear model against fitted polynomials, that when the 'true' model fits well, it is rejected about the same number of times as it should be according to the chosen level of significance; and even when it does not fit well, the number of rejections of the 'true' hypothesis is only of the order of 25-30% in excess of the number that should be rejected. Insofar as these results are generally applicable, it would seem that this version of the Hoel-test is a useful device for the purposes of data analysis.

(ii) *Numerical Methods of Identification*

When the variances are equal but θ_0 and θ_1 are unknown, λ is not identifiable in the comprehensive model (5.2.2). A practical way around the lack of identifiability of λ in the comprehensive model is to use a consistent estimate of at least one of θ_0 and θ_1 under their respective hypotheses. This procedure is not, of course, equivalent to imposing identifying restrictions in the usual sense, but it does enable calculation of an estimate of λ. The estimate of λ obtained has no meaning in the comprehensive model simply because it is not identifiable in that model. Therefore, when arbitrary numerical methods are used as 'identifying' restrictions, the 'unrestricted' model reduces to an algorithm, since it cannot reasonably be 'accepted'. However, the t-ratio for $\hat{\lambda}$ turns out to be a valid statistic for testing H_0, and is asymptotically distributed as $N(0,1)$ under H_0.

Davidson and MacKinnon (1980) have developed several tests of the kind just described for both linear and non-linear models. Consider the linear case, where equation (5.2.2) is rewritten:

$$(5.2.4) \quad y = (1-\lambda)X\theta_0 + \lambda Z\theta_1 + u$$

and it is clear that λ is not identifiable. It is assumed that the columns of X and Z are linearly independent but not orthogonal. Replace θ_1 in (5.2.4) by its least squares estimator under H_1, leading to:

$$(5.2.5) \quad y = (1-\lambda)X\theta_0 + \lambda P_z y + u$$

and use the t-ratio for $\hat{\lambda}$, called the J-statistic, to test H_0. Equation (5.2.5) is noteworthy since its non-linear equivalent is the starting point for the analysis of Davidson and MacKinnon (1980). Thus, for these authors, the artificial model (5.2.5) is simply an algorithm and does not arise by virtue of nesting of the component likelihood functions. A more helpful way of looking at (5.2.5), rather than by way of definition, is to regard it as a linear combination of the linear counterparts of (5.1) and (5.2) with weights $(1-\lambda)$ and λ, respectively, where $u = \lambda Z(\theta_1 - \hat{\theta}_1) + (1-\lambda) u_0 + \lambda u_1$.

The process of replacing θ_1 in (5.2.4) by $\hat{\theta}_1$ is arbitrary, since θ_1 may be replaced by *any* estimate that is asymptotically uncorrelated with the disturbance under H_0. The suggestion of Atkinson (1970) may also be applied to the J-test. For practical purposes, replace $Z\theta_1$ with the estimate of the expected value of $P_z y$ under H_0, namely $P_z P_x y$. Equation (5.2.5) becomes

$$(5.2.6) \quad y = (1-\lambda)X\theta_0 + \lambda P_z P_x y + u \quad .$$

Assuming that $(1/n)X^T X$, $(1/n)Z^T Z$ and $(1/n)X^T Z$ converge to well-defined finite matrices, the first two positive-definite and the third non-zero, it is straight-forward to show that, under H_0, the t-ratio for $\hat{\lambda}$ of (5.2.6) is asymptotically equivalent to the J- statistic from (5.2.5). Since there are, in principle, an

infinite number of linear estimators, there must be an infinite number of t-ratios that are appropriate for testing H_0. Notice carefully that the estimate of λ is of no interest in itself for testing H_0, just as the estimate of δ was of no interest in (4.2). Since $Z\theta_1$ in (5.2.4) is replaced by an arbitrary consistent estimate, $\hat{\lambda}$ is itself arbitrary. In practical applications of the AN-principle, it is obvious that an *infinite* number of methods exist. While the arbitrariness of the substitution may be inconsequential in large samples (see McAleer, 1980a), a more relevant question is whether the route *via* $\hat{\theta}_1$ is the 'best' way to proceed.

5.3 *The Minimum Error Variance Criterion*

A study of the principles and methods available for testing non-nested hypotheses would be incomplete without mentioning the Minimum Error Variance (MEV) criterion. Suppose there are two competing non-nested, (possibly) non-linear regression models, one of which

$$(5.3.1) \quad H_0: \quad y = f(\theta_0;X) + u_0 \quad , \quad u_0 \sim N(0, I_n\sigma_0^2)$$

is 'true', the other

$$(5.3.2) \quad H_1: \quad y = g(\theta_1;Z) + u_1$$

being false. The way in which H_1 is false is not specified; in particular, no assumption is made concerning u_1. A model selection criterion examined by Theil (1961, p. 213) for the linear case says that if one of the models being considered is 'true', in the above sense, then

$$(5.3.3) \quad E_0(s_1^2) = E_0(s_0^2) + (n-k_1)^{-1} \theta_0'X' (I-P_z)X\theta_0 \geqq E_0(s_0^2)$$

where E_0 denotes expected value under H_0, s_0^2 and s_1^2 are the standard least

squares estimates of the error variances under H_0 and H_1, respectively, and $E_0(s_0^2) = \sigma_0^2$. Since the expected value of the estimated error variance under the 'true' model will never be larger than under the false model, Theil suggests that choosing the model with the minimum estimated error variance will lead to the selection of the correct model, 'on average'. Kloek (1975) has shown that, under fairly general conditions and in large samples, the MEV criterion will choose the 'true' model with probability arbitrarily close to unity.

The basis MEV criterion applies to non-nested linear regression models with non-stochastic regressors, when one of the models considered is 'true'. This result has been extended by Schmidt (1974) to allow for autocorrelated disturbances, and by Giles and Smith (1977) to the case of linear restrictions. The former also noted that the rule holds asymptotically under ordinary least squares estimation if the regressors are stochastic. McAleer (1980b) has extended Theil's MEV rule by providing an asymptotic justification for choosing between non-nested, non-linear regression models.

The device of selecting the model with minimum estimated error variance certainly has a practical appeal. However, its viability depends upon the requirement that one of the models being considered is 'true'. From an econometric viewpoint, an important criticism of the MEV approach is that the term model 'selection', by its very nature, ensures that one model will inevitably be chosen. It is not possible, then, for both models to be either rejected or 'accepted' by the data. The problem of model *selection* might more reasonably be viewed as one of model *testing* (see Sections 5.1 and 5.2 above), but the decision to use one approach or the other naturally depends upon the specific interests of the researcher. Be that as it may, the MEV rule is a useful procedure when it is intended to choose between alternative non-nested regression

models. For the MEV criterion, a method is certainly involved but there is neither an artificial model nor the use of a modified likelihood ratio. The decision to choose a particular model on the basis of minimum estimated error variance is an arbitrary rule, so that no principle of testing is involved here.

6. CONCLUDING REMARKS AND RESEARCH PERSPECTIVES

In this paper we have attempted to synthesize the principles on which testing is to be based and the practical statistical methods which they generate. We illustrated the general rules which specify the distinction in the case of nested hypotheses with several examples. The concept of an artificial model was examined in testing for misspecification. Two principles were outlined for testing non-nested hypotheses. The Principle of Artificial Nesting was shown to accommodate artificial models that were of interest in their own right, in that they could be 'acceptable' to the data, and artificial models that were mere algorithms. In the former case, a test based upon the artificial model may be regarded as conforming to the W-principle in that it permits 'accepting' an 'unrestricted' form, whereas in the latter no 'unrestricted' form is tolerated. It is then more natural to appeal to the M-principle as a basis for testing.

Although the exponential weighting of likelihood functions considered in Section 5 is naturally appealing, there are certainly other methods available for nesting alternative models within a more general framework. This raises the interesting question of whether the concept of non-nested hypotheses is itself more than a thoroughly practical artifact. It could be argued, for example, that reasonable competing hypotheses are just special cases of some more general system, special components of which are the individual hypotheses H_0 and H_1.

A more general method of combining two models, given equality of their respective variances, is to form the weighted mean of order ρ. Pesaran (1979) suggested nesting two likelihood functions using this approach but did not derive the implicit artificial regression model. Another procedure is to construct the weighted mean of order ρ of the deterministic components of the two hypotheses. Recalling that x_i and z_i are the i'th rows of X and Z, respectively, and $P_z y = Z\hat{\theta}_1$, the *non-linear* nesting scheme for the i'th observation is *defined* as

$$(6.1) \qquad y_i = \{(1-\lambda)(x_i\theta_0)^\rho + \lambda(z_i\hat{\theta}_1)^\rho\}^{1/\rho} + u_i \quad , \quad i=1,2,\ldots,n$$

where it is assumed that $x_i\theta_0$, $x_i\hat{\theta}_0$, $z_i\theta_1$ and $z_i\hat{\theta}_1$ are positive for all i and that $u_i \sim NID(0, \sigma^2)$. Equation (6.1) simplifies to the i'th equation of (5.2.5) upon setting $\rho=1$, and to

$$(6.2) \qquad y_i = (x_i\theta_0)^{1-\lambda} (z_i\hat{\theta}_1)^\lambda + u_i$$

upon letting $\rho \to 0$. A possible advantage of (6.1) over both (5.2.5) and (6.2) is that there is a nesting parameter ρ, as well as a testing paramater λ. These may in principle be estimated separately, thereby enabling a test of the *form* of nesting as well as a test of the non-nested hypotheses themselves. This should be distinguished from (5.2.3), where the testing parameter is identifiable but the nesting parameter is not, and from (5.2.2), where the nesting and the testing parameters are identical. But note carefully that (6.1) does not arise by virtue of combining the likelihood functions of the two models as expressed in equations (5.1) and (5.2). Rather, we arbitrarily *define* the general model (6.1) which, upon restriction of λ at zero, yields H_0; and similarly for H_1 when the roles are reversed. For a comprehensive discussion of the various forms of nesting and their consequences, see McAleer (1980a). Consideration of the form of nesting may turn out to be as important as the actual testing procedures themselves.

In respect of prior information, the analysis could reasonably be conducted in terms of small samples. When we rely on numerical methods instead, an infinite number of estimators, and hence an infinite number of asymptotic test statistics, become available. For example, it was shown to be easy to apply Atkinson's (1970) modification of the Cox (1961, 1962) procedure to a test of Davidson and MacKinnon (1980). In view of the plethora of large-sample tests that are available, an analysis of their small-sample properties is obviously an important avenue for future research.

REFERENCES

Aitchison, J. and S.D. Silvey (1958) "Maximum-likelihood Estimation of Parameters Subject to Restraints," *Annals of Mathematical Statistics*, 29, 813-828.

Anderson, T.W. and H. Rubin (1949) "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20, 46-63.

Anderson, T.W. and H. Rubin (1950) "The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 21, 570-582.

Atkinson, A.C. (1970) "A Method for Discriminating Between Models," *Journal of the Royal Statistical Society (Series B)*, 32, 323-344.

Basmann, R.L. (1960) "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics," *Journal of the American Statistical Association*, 55, 650-659.

Breusch, T.S. and A.R. Pagan (1980) "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics," *Review of Economic Studies*, 47, 239-253.

Brown, R.L., J. Durbin, and J.M. Evans (1975) "Techniques for Testing the Constancy of Regression Relationships over Time," *Journal of the Royal Statistical Society (Series B)*, 37, 149-192.

Chow, G.C. (1960) "Tests of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28, 591-605.

Cox, D.R. (1961) "Tests of Separate Families of Hypotheses," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1 (Berkeley: University of California Press), 105-123.

Cox, D.R. (1962) "Further Results on Tests of Separate Families of Hypotheses," *Journal of the Royal Statistical Society (Series B)*, 24, 406-424.

Davidson, R. and J. MacKinnon (1980) "Several Tests for Model Specification in the Presence of Alternative Hypotheses," forthcoming *Econometrica*.

Dhrymes, P.J. (1968) "Alternative Asymptotic Tests of Significance and Related Aspects of 2SLS and 3SLS Estimated Parameters," *Review of Economic Studies*, 33, 213-226.

Durbin, J. (1954) "Errors in Variables," *Review of the International Statistical Institute*, 22, 23-32.

Fisher, G.R. (1977) "The Linear Hypothesis and Large Sample Theory: Extension to the Case of Endogenous Variables," Paper prepared for Session M30, European Meeting of the Econometric Society, Vienna, Austria.

Fisher, G.R. (1978) "The Linear Hypothesis and Structural Change," Unpublished paper, Department of Economics, Queen's University, February 1978.

Fisher, G.R. (1980) "Two Types of Residuals and the Classical Identifiability Test Statistic," forthcoming in *Proceedings of the Econometric Society European Meeting '79 - Selected Econometric Papers in Memory of Stefan Valavanis* (North Holland).

Fisher, G.R. and M. McAleer (1980a) "On the Interpretation of the Cox Test in Econometrics," forthcoming in *Economics Letters*.

Fisher, G.R. and M. McAleer (1980b) "Alternative Procedures and Associated Tests of Significance for Non-nested Hypotheses," Invited paper to the Conference on Model Selection, Gainesville, Florida, April 1980.

Giles, D.E.A. and R.G. Smith (1977) "A Note on the Minimum Error Variance Rule and the Restricted Regression Model", *International Economic Review*, 18, 247-251.

Hausman, J.A. (1978) "Specification Tests in Econometrics," *Econometrica*, 46, 1251-1271.

Hoel, P.G. (1947) "On the Choice of Forecasting Formulas," *Journal of the American Statistical Association*, 42, 605-611.

Kloek, T. (1975) "Note on a Large-Sample Result in Specification Analysis," *Econometrica*, 43, 933-936.

McAleer, M. (1980a) "On Two Classes of Model Testing Procedures," Unpublished paper, Department of Economics, Queen's University, March 1980.

McAleer, M. (1980b) "The Minimum Error Variance Rule for Non-linear Regression Models," Unpublished paper, Department of Economics, Queen's University, July 1980.

Morgan, A. and W. Vandaele (1974) "On Testing Hypotheses in Simultaneous Equation Models," *Journal of Econometrics*, 2, 55-65.

Pereira, B. de B. (1977) "A Note on the Consistency and on the Finite Sample Comparisons of Some Tests of Separate Families of Hypotheses," *Biometrica*, 64, 108-114.

Pesaran, M.H. (1974) "On the General Problem of Model Selection," *Review of Economic Studies*, 41, 153-171.

Pesaran, M.H. (1979) "A Note on the Comprehensive Method of Testing Non-linear and Non-nested Regression Models," Unpublished paper, Trinity College, University of Cambridge, July 1979.

Pesaran, M.H. and A.S. Deaton (1978) "Testing Non-nested Non-linear Regression Models," *Econometrica*, 46, 677-694.

Quandt, R.E. (1974) "A Comparison of Methods for Testing Non-nested Hypotheses," *Review of Economics and Statistics*, 56, 92-99.

Rao, C.R. (1948) "Tests of Significance in Multivariate Analysis," *Biometrika*, 35, 58-79.

Schmidt, P. (1974) "A Note on Theil's Minimum Standard Error Criterion When the Disturbances are Autocorrelated," *Review of Economics and Statistics*, 56, 122-123.

Seber, G.A.F. (1964) "The Linear Hypothesis and Idempotent Matrices," *Journal of the Royal Statistical Society (Series B)*, 26, 261-266.

Silvey, S.D. (1959) "The Lagrange-multiplier Test," *Annals of Mathematical Statistics*, 30, 389-407.

Theil, H. (1961) *Economic Forecasts and Policy* (Amsterdam: North Holland).

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.

Weatherburn, C.E. (1952) *A First Course in Mathematical Statistics* (Cambridge, Cambridge University Press).

White, H. (1980) "Maximum Likelihood Estimation of Misspecified Models: II," Discussion Paper 80-2, Department of Economics, University of Rochester, January 1980.

Wilks, S.S. (1961) *Mathematical Statistics* (New York, Wiley).