ISSN 0316-5078

# INSTITUTE FOR ECONOMIC RESEARCH

QUEEN'S UNIVERSITY



Kingston, Ontario, Canada K7L 3N6

# PROPOSITIONS, PRINCIPLES AND METHODS: THE LINEAR HYPOTHESIS AND STRUCTURAL CHANGE*

Gordon Fisher

Department of Economics
Queen's University

Discussion Paper No. 410

June 1980

*An earlier version of the first three sections of this paper appeared in March 1980 as Discussion Paper No. 377 under the title: "Propositions, Principles and Methods: The Case of the Linear Hypothesis". Section 4 has now been added. This originally appeared as a Working Paper in 1978 under the title: "The Linear Hypothesis and Structural Change".

# PROPOSITIONS, PRINCIPLES AND METHODS: THE LINEAR HYPOTHESIS AND STRUCTURAL CHANGE

*ABSTRACT*

This paper seeks to distinguish between the principles upon which testing of statistical hypotheses may be based and the practical methods which these principles generate. 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 the principles. Simple relations between various test statistics and their distributions are outlined. These are then applied to two well-known methods of testing for structural change. It is found that the cusum of squares test using recursive residuals is based upon the Lagrange Multiplier Principle while the prediction interval (or Chow-) test is based upon the Wald Principle. Relations between these tests and some straightforward extensions are outlined.

Gordon Fisher
June, 1980

# PROPOSITIONS, PRINCIPLES AND METHODS: THE LINEAR HYPOTHESIS AND STRUCTURAL CHANGE

- Gordon Fisher -

## 1. *PREAMBLE*

The testing of propositions put forward hypothetically is fundamental to the advancement of knowledge in science, because it permits the classification of potentially fruitful lines of enquiry into those that are worth pursuing and those that are not. Unfortunately, the outcome of a valid testing procedure may be disputed because, *inter alia*, it comprises an arbitrary element, and hence it is possible for two admissible tests of the same kind to yield conflicting outcomes. For these reasons, testing may well raise more questions than it answers and the way forward may not be clear. Characteristics of this kind are especially true of testing statistical hypotheses. Consequently, it is well to have clear in our minds, before such a test is applied, what its special characteristics might be and how, in view of these, it might perform relative to some alternative test. In this respect, it is helpful to distinguish the principles on which testing is to be based from the practical statistical methods which these principles generate. By a *principle* here is meant a general rule which specifies how tests are to be devised. By a *method* is meant 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 consise in terms of principles than in terms of methods, since the former avoids unhelpful repetition of notation and ideas. Moreoever, knowledge that different methods have a common root in a particular principle is a useful aid to memory.

## 2. *PRINCIPLES AND METHODS*

The distinction to be drawn between principles and methods 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 $\theta$ from a random sample of n observations from a given distribution; $\theta$ is unknown, save that it lies in p-dimensional Euclidean space $\Omega (p<n)$. It is desired to

test $H_0$: $\theta \in \omega$, where $\omega$ represents a sub-set of points in $\Omega$ which obey the $r<p$ restrictions $h(\theta) = 0$. If $\theta_\Omega$ denotes maximum-likelihood estimate of $\theta$ in $\Omega$ (*i.e. unrestricted* maximum-likelihood estimation), then the Wald (W-) test for $H_0$ is given by

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

a standardized quadratic form in $h(\theta_\Omega)$, where $D_\Omega\{\cdot\}$ denotes dispersion matrix corresponding to (unrestricted maximum-likelihood) estimation in $\Omega$. Subject to regularity, $n^{-\frac{1}{2}}h(\theta_\Omega)$ is distributed as $N(0, H_\theta^T B_\theta^{-1} H_\theta)$ asymptotically, where $H_\theta$ is the matrix of partial derivatives $\partial h^T(\theta)/\partial\theta$ and $B_\theta$ is the information matrix corresponding to a single observation, both being evaluated at the true value of $\theta$. Consequently,

$$(2.2) \qquad D_\Omega\{h(\theta_\Omega)\} = n^{-1} H_\theta^T B_\theta^{-1} H_\theta \text{ at } \theta = \theta_\Omega$$

and hence

$$(2.3) \qquad W = nh^T(\theta_\Omega)[H^T_{\theta_\Omega} B^{-1}_{\theta_\Omega} H_{\theta_\Omega}]^{-1} h(\theta_\Omega)$$

is approximately distributed as $\chi^2(r)$ under $H_0$.

Similarly, if $\theta_\Omega$ now refers to another asymptotic normal estimator of $\theta$ in $\Omega$ 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 $\theta_\Omega$ is unbiased and exactly normal, then a small-sample refinement of W based on the F-distribution may be obtained. I will return to this below.

Notice that, whatever W-test is used, its associated estimates are invariably based upon unrestricted estimation; that is, upon estimation of $\theta$ in $\Omega$, 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 $\theta$ in $\omega$; that is, upon *restricted* estimation, using $\phi_\omega$, the estimate of the vector-valued Lagrange multiplier corresponding to $h(\theta) = 0$. The large-sample test based upon the M-principle is given by

$$(2.4) \qquad M = \phi_\omega^T [D_\omega(\phi_\omega)]^{-1} \phi_\omega \quad ,$$

where $D_\omega(\cdot)$ denotes dispersion matrix corresponding to (restricted) estimation in $\omega$. For example, if $L(y,\theta)$ denotes the value of the likelihood function corresponding to a given observed n-sample y at the point $\theta$ then the (restricted) maximum-likelihood estimates must satisfy:

$$D_\theta \log L(y,\theta_\omega) - H_{\theta_\omega} \phi_\omega = 0$$

$$h(\theta_\omega) = 0$$

and the first of these implies $D_\theta \log L(y,\theta_\omega) = H_{\theta_\omega} \phi_\omega$. Moreover, it may be shown that, subject to regularity, $n^{-\frac{1}{2}}\phi_\omega$ tends in distribution to $N(0,\{H_\theta^T B_\theta^{-1} H_\theta\}^{-1})$, and hence that

$$(2.5) \qquad M = n^{-1}\phi_\omega^T H_{\theta_\omega}^T B_{\theta_\omega}^{-1} H_{\theta_\omega} \phi_\omega = n^{-1}\{D_\theta \log L(y,\theta_\omega)\}^T B_{\theta_\omega}^{-1}\{D_\theta \log L(y,\theta_\omega)\}$$

is approximately distributed as $\chi^2(r)$ under $H_0$. The last expression in (2.5) is the score test (Rao, 1948). Incidentally, there is no need to insist on maximum-likelihood estimation: the estimated Lagrange-multiplier $\phi_\omega$ 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. *APPLICATION*

We shall now consider a particular application of the principles introduced in Section 2. 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 $\mu$ and $\sigma^2$ are unknown. Corresponding to the sub-space $\Omega$, the least squares estimates of $\mu$ and $\sigma^2$ are denoted by $m_\Omega$ and $s^2_\Omega$ respectively. It is desired to test the linear hypothesis $H_0$: $\mu \in \omega$; $\omega \subset \Omega$, where $\omega$ is (p-r)-dimensional. The number r represents the number of linear restrictions on $\Omega$ to define $\omega$. Least squares estimation under $H_0$ yields $m_\omega$ and $s_\omega^2$.

The standard test-statistic for $H_0$ is:

$$(3.1)\qquad F = \frac{y^T(P_\Omega - P_\omega)y}{y^T(I_n - P_\Omega)y} \cdot \frac{n-p}{r}$$

where P denotes an orthogonal projection: $P_\Omega$ is on $\Omega$ along $\Omega^\perp$ and $P_\omega$ is on $\omega$ along $\omega^\perp$, orthogonal complementation ($\perp$) being relative to $\mathcal{E}_n$. More explicitly, if $\omega$ is defined by $\omega \equiv \Omega \cap N[A^T]$, where $A^T$ is a known r x n matrix of rank $r \leqq p$, then any $x \in \Omega$ which obeys $A^Tx = 0$ must lie in $\omega$. 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:

$$(3.2)\qquad P_\Omega - P_\omega = P_\Omega A(A^TP_\Omega A)^{-1}A^TP_\Omega \quad ,$$

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) embodies the W-principle since $m_\Omega = P_\Omega y$ and hence

$$(3.3)\qquad F = \frac{y^T(P_\Omega - P_\omega)y}{rs_\Omega^2} = \frac{(A^Tm_\Omega)^T[D_\Omega(A^Tm_\Omega)]^{-1}(A^Tm_\Omega)}{r}$$

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) and (3.3) each have the central F(r, n-p) distribution under $H_0$. Further, rF is a quadratic form based upon the unrestricted estimates $m_\Omega$ and $s_\Omega^2$ and the given restrictions only; upon replacing $s_\Omega^2$ with $\sigma^2$, it is seen to be a quadratic form in standardized normal variates, exactly under $H_0$. Since also $s^2_\Omega$ is asymptotically equivalent to $\sigma^2_\Omega$, the maximum-likelihood estimator of $\sigma^2$, it is obvious that $rF \sim \chi^2(r)$ for large n. Indeed, if $s^2_\Omega$ is replaced by $\sigma^2_\Omega$, we may write $rF = W$ to comply with the original definition of W.

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

$$(3.4) \qquad L = (y-\mu)^T(y-\mu) + 2\mu^T A\phi - 2\mu^T(I-P_\Omega)\kappa$$

for variations in $\mu$ and the vector Lagrange multipliers $\phi$ and $\kappa$ 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 $\phi$ from (3.4) and the implicit hypothesis corresponding to $H_0$, namely: $\phi = 0$. Note carefully that the entire procedure is based upon least squares estimation of $\mu$ in $\omega$. Writing $f_\omega$ for the estimate of $\phi$ corresponding to $m_\omega$, the first-order conditions from (3.4) lead to:

$$\begin{bmatrix} I_n & P_\Omega A \\ A^T P_\Omega & 0 \end{bmatrix} \begin{bmatrix} m_\omega \\ f_\omega \end{bmatrix} = \begin{bmatrix} P_\Omega y \\ 0 \end{bmatrix} .$$

This yields:

$$\begin{bmatrix} m_\omega \\ f_\omega \end{bmatrix} = \begin{bmatrix} I_n - P_\Omega A(A^T P_\Omega A)^{-1} A^T P_\Omega & P_\Omega A(A^T P_\Omega A)^{-1} \\ (A^T P_\Omega A)^{-1} A^T P_\Omega & -(A^T P_\Omega A)^{-1} \end{bmatrix} \begin{bmatrix} P_\Omega y \\ 0 \end{bmatrix} ;$$

in particular, $f_\omega = (A^T P_\Omega A)^{-1} A^T P_\Omega y$ and $m_\omega = P_\Omega y - P_\Omega A(A^T P_\Omega A)^{-1} A^T P_\Omega y$. Corresponding to these estimates

$$(3.6) \qquad s_\omega^2 = \frac{y^T(I_n - P_\omega)y}{n-p+r} = \frac{(y-m_\omega)^T(y-m_\omega)}{n-p+r}$$

is an unbiased estimate of $\sigma^2$ under $H_0$. The small-sample test statistic based upon the M-principle uses only estimates corresponding to estimation of $\mu$ in $\omega$. If the statistic is M, then

$$M = f_\omega^T\left[D_\omega(f_\omega)\right]^{-1} f_\omega$$

or

$$(3.7) \qquad M = y^T P_\Omega A(A^T P_\Omega A)^{-1} \left[(A^T P_\Omega A)^{-1} s_\omega^2\right]^{-1} (A^T P_\Omega A)^{-1} A^T P_\Omega y \quad .$$

Seber's (1964) demonstration that both the W- and the M-principles lead to exactly the same test statistic lies in noting that, if $\sigma_\Omega^2$, the unrestricted maximum-likelihood estimate of $\sigma^2$, replaces $s_\omega^2$ in (3.7), then the resulting expression is $\{nr/(n-p)\}$ times the F-statistic of (3.1). Unfortunately, this step involves evaluating the dispersion matrix of $f_\omega$, namely $(A^T P_\Omega A)^{-1}\sigma^2$, at the point $\sigma^2 = \sigma_\Omega^2$, that is, at an *unrestricted* estimate. Clearly this violates the M-principle as established above, since this principle requires the use of restricted estimates only. Moreover, using $\sigma_\Omega^2$ in place of $s_\omega^2$ is unnecessary to demonstrate that the two principles lead to the same F-statistic, as the following argument reveals.

Equation (3.7) is readily seen to reduce to

$$(3.8) \qquad M = \frac{y^T(P_\Omega - P_\omega)y}{y^T(I_n - P_\omega)y} \cdot (n-p+r)$$

upon application of (3.2). Under $H_0$, $\{M/(n-p+r)\}$ is distributed as $\beta_1(\frac{r}{2}, \frac{n-p}{2})$ exactly, since

$$(3.9) \qquad \frac{y^T(P_\Omega - P_\omega)y}{y^T(I_n - P_\omega)y} = \frac{y^T(P_\Omega - P_\omega)y}{y^T(P_\Omega - P_\omega)y + y^T(I_n - P_\Omega)y}$$

and the two components in the denominator of the right-hand side of (3.9), each divided by $\sigma^2$, 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) and (3.3), and M, in equation (3.8), may be written

$$(3.10) \qquad \frac{W}{n-p + W} = \frac{M}{n-p+r}$$

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

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

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

$$(3.11) \qquad \lambda^{\frac{2}{n}} = \frac{M}{W} \cdot \frac{n-p}{n-p+r} \quad .$$

Note also that

$$(3.12) \qquad W = (\lambda^{\frac{2}{n}}-1)(n-p) \quad .$$

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 (3.9), (3.10) and (3.12) that

$$(3.13) \qquad W = (\lambda^{\frac{2}{n}}-1)(n-p) \geqq M\{(n-p)/(n-p+r)\}$$

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

$$(3.14) \qquad W\{\frac{n}{n-p}\} \geqq \{-2 \log \lambda\} \geqq M\{\frac{n}{n-p+r}\} \quad ,$$

each of which has the $\chi^2(r)$-distribution for large n.

Relations (3.10) - (3.12) 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 $\lambda$ 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).

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 $\sigma^2$ that is, or should be, used. In the former case, an estimate of $\sigma^2$ in $\Omega$ is required, for example $\sigma^2_\Omega$ or $s^2_\Omega$; when the M-principle is applied, it is necessary to use an estimate of $\sigma^2$ in $\omega$.

## 4. *AN EXAMPLE*

It is common in applications of regression analysis that estimates of the coefficients have been obtained on the basis of m observations and another (n-m) observations become available. It is then important to enquire whether, on the basis of the data, the extra (n-m) observations may be deemed to have been generated by the same process that generated the initial m. Various tests have been devised to deal with this, and similar, situations. If the regression is linear with $p<m$ unknown coefficients and has independent, homoskedastic, normal errors then, given $(n-m)>p$, the analysis of covariance is an appropriate procedure. If $(n-m)\leqq p$, the prediction interval test may be used. Essentially, this has exactly the same linear structure as the analysis of covariance, except the latter has to be modified to deal with the singularity introduced by the condition $(n-m)\leqq p$. The origins of this test are a little obscure, but Chow (1960) was certainly the first to apply the test specifically to the idea of structural change in economics. Another test, called the cusum of squares test, using recursive residuals, was introduced by Brown *et al.* (1975) to help examine the general question of the constancy of regression relationships over time. Harvey (1976) has discussed the prediction interval (or Chow-) test in terms of recursive residuals and has extended its application to sequences of blocks of observations. 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.

It will be helpful to begin by regarding each of the extra (n-m) observations as generated by a new process. Let $y_t$ be an observation on the dependent variable at time t and $\varepsilon_t$ a zero mean regression error at the same time. Each of these is normally distributed, independently over t, with common variance $\sigma^2$. The regression relationship is written:

$$(4.1) \qquad y_t = x_t^T\beta_t + \varepsilon_t \quad , \quad t = 1,2,\dots,m,m+1,\dots,n,$$

there being p components in the vectors $x_t$ and $\beta_t$, the latter being unknown save that $\beta_t = \beta_m$ for $t = 1,2,\dots,m$. It is presumed that the $x_t$ are fixed in repeated samples. This formulation may also be written:

$$(4.2) \qquad \begin{bmatrix} y_m \\ y_{m+1} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} X_m & 0 & 0 & \dots & 0 \\ 0 & x_{m+1}^T & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & x_n^T \end{bmatrix} \begin{bmatrix} \beta_m \\ \beta_{m+1} \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_m \\ \varepsilon_{m+1} \\ \vdots \\ \varepsilon_n \end{bmatrix} .$$

In equation (4.2), $y_m$ and $\varepsilon_m$ are m x 1 vectors, corresponding to the first m observations, otherwise $y_t$ and $\varepsilon_t$ are scalar-valued as in (4.1); $X_q$ (q=m) denotes the (fixed) matrix of rows $x_t^T$, $t = 1,2,\dots,q$.

The least squares estimate of $\beta_m$ in (4.2) may be found; but $\beta_{m+1},\beta_{m+2},\dots,\beta_n$ are not estimable *via* least squares because each is associated with one observation only on each of y and x. Thus, for these observations, it is always possible to 'estimate' $y_t$ without error, by appropriate choice of $\beta_t$. To accommodate this detail, the non-singular matrix $\hat{X}_q$ is introduced.

$$(4.3)\qquad \hat{X}_q = \begin{bmatrix} X_q & 0 \\ 0 & I_{n-q} \end{bmatrix}, \quad q = p+1, p+2, \ldots, m, m+1, \ldots, n$$

where $\hat{X}_n = X_n$. The least squares estimate of the mean of y in (4.2) may be regarded as lying in the span of $\hat{X}_q$ when q=m, since $\beta_{m+1}, \beta_{m+2}, \ldots, \beta_n$ are essentially arbitrary in this case, and hence may be taken as $y_t$ itself for t = m+1,m+2,...,n. An additional advantage in using $\hat{X}_q$ is that the sequence of hypotheses: $\beta_m = \beta_{m+1}$; $\beta_m = \beta_{m+1} = \beta_{m+2}$; and so on, may each be associated with a corresponding $\hat{X}_q$ for successive values of q. The perpendicular projection on the span of $\hat{X}_q$ is $\hat{P}_q$ which takes the form:

$$(4.4)\qquad \hat{P}_q = \begin{bmatrix} X_q(X_q^T X_q)^{-1} X_q^T & 0 \\ 0 & I_{n-q} \end{bmatrix} = \hat{X}_q(\hat{X}_q^T \hat{X}_q)^{-1} \hat{X}_q^T \quad .$$

In this setting, suppose it is desired to test the hypothesis $H_0 = \beta_t = \beta$ for all t, against the alternative $H_\alpha$: $\beta_t = \beta_I$ for t = 1,2,...,m; $\beta_t = \beta_{II}$ for t = m+1,m+2,...,n, that is, the null hypothesis that all the β's are the same, against the alternative that there is one value up to t=m and another after that, from t = m+1 to t = n. Assuming $(n-m) \leqq p$, the prediction interval (or Chow-) test is

$$(4.5)\qquad F = \frac{y^T(\hat{P}_m - \hat{P}_n)y}{y^T(I_n - \hat{P}_m)y} \cdot \frac{m-p}{n-m}$$

which has the central F(n-m, m-p) distribution under $H_0$. Clearly F in (4.5) is based upon the W-principle since the denominator divided by (m-p) is the unbiased least squares estimate of $\sigma^2$ from the 'unrestricted regression' (4.2), namely the regression corresponding to the first m observations. In the restricted case

$$(4.6) \qquad y = \hat{X}_n \beta + \varepsilon \quad ,$$

and so the restricted estimate of $\sigma^2$ is $\{(n-p)^{-1}\}\{y^T(I_n-\hat{P}_n)y\}$. Thus the test corresponding to the M-principle is

$$(4.7) \qquad M = \frac{y^T(\hat{P}_m-\hat{P}_n)y}{y^T(I_n-\hat{P}_n)y} \cdot (n-p)$$

where $M/(n-p)$ has the $\beta_1(\frac{n-m}{2}, \frac{m-p}{2})$ distribution under $H_0$. In terms of earlier notation, $\Omega$ is the span of $\hat{X}_m$ and $\omega$ is the span of $\hat{X}_n$. Moreover, (4.5) corresponds with (3.1), and (4.7) with (3.8). Of course, if $(n-m)>p$, it is more appropriate to apply the analysis of covariance in the sense that this test is never less powerful than the prediction interval test (see *e.g.* Saw, 1964). However, this is a mere detail since all that is involved is a re-definition of $\hat{P}_m$ in an obvious way: namely, replace the sub-matrix $I_{n-m}$ in $\hat{P}_m$ by the perpendicular projection on the span of the matrix comprising $(n-m)$ rows $x_t^T$, $t = m+1, m+2, \ldots, n$.

As far as recursive residuals are concerned, if the first m observations alone are considered, there are $(m-p)$ recursive residuals $u_q$, $q = p+1, p+2, \ldots, m$, defined by:

$$(4.8) \qquad u_q = \frac{y_q - x_q^T b_{q-1}}{(1-x_q^T(X_{q-1}^T X_{q-1})^{-1} x_q)^{\frac{1}{2}}} \quad ,$$

where $b_q$ is the least squares estimate of $\beta_m$ based upon the first q observations. If all n observations are considered, the upper bound on q may be extended to include all the extra $(n-m)$ observations. An alternative expression for $u_q$ is:

$$u_q^2 = y^T(I_n-\hat{P}_q)y - y^T(I_n-\hat{P}_{q-1})y$$

or

$$(4.9) \qquad u_q^2 = y^T(\hat{P}_{q-1}-\hat{P}_q)y, \qquad q = p+1,p+2,\ldots,m,m+1,\ldots,n.$$

Thus $\Sigma_{p+1}^m u_q^2 = y^T(I_n-\hat{P}_m)y$, *i.e.* the residual sum of squares from the regression on the first m observations, since $\hat{P}_p = I_n$. For proofs of these results see Phillips and Harvey (1975) and Brown *et al.* (1975). Equation (4.5) may be re-written as

$$(4.10) \qquad F = \frac{\Sigma_{m+1}^n u_q^2/(n-m)}{\Sigma_{p+1}^m u_q^2/(m-p)}$$

and equation (4.7) as

$$(4.11) \qquad M = \frac{\Sigma_{m+1}^n u_q^2}{\Sigma_{p+1}^n u_q^2} \cdot (n-p) \quad .$$

In the above examples, there are presumed to be good *a priori* reasons for asking whether the extra (n-m) observations belong to the same regime as the first m. The alternative hypothesis in this case requires $\beta_{m+1}=\beta_{m+2}=\ldots=\beta_n$, but to be different from $\beta_I$ (the coefficient for the first m observations) within the specification (4.2). In other cases, the form of the change may not be specified, and the point (or points) at which change has taken place need not be known. In cases of this kind the alternative hypothesis is left deliberately more vague than it is in respect of the tests examined in (4.5), (4.7), (4.10) and (4.11); indeed as vague as is displayed in equation (4.2) itself. When such vagueness is in order, the cusum of squares test, as outlined in Brown *et al.* (1975), may be used. Thus, if it is believed that structural changes may have occurred within the total of n sample points available at any point s = p+1,p+2,...,n, then the test statistic is $Q_s$ where:

$$(4.12)\qquad Q_s = \frac{\Sigma_{p+1}^{s} u_q^2}{\Sigma_{p+1}^{n} u_q^2} = \frac{y^T(I_n-\hat{P}_s)y}{y^T(I_n-\hat{P}_n)y} ,$$

which is type 1 beta distributed with parameters $\{(s-p)/2\}$ and $\{(n-s)/2\}$, under the null hypothesis that all the $\beta_t$'s in (4.1) are equal. Comparing (4.12) with (4.11) and (4.7), $(1-Q_s)$ is exactly $\{M/(n-p)\}$ with $m=s$, which may be recognized by writing $M_s = (n-p)(1-Q_s)$. In this sense, $Q_s$ may be regarded as based upon the M-principle. Moreover,

$$(4.13)\qquad \frac{(1-Q_s)/(n-s)}{Q_s/(s-p)} = \frac{y^T(\hat{P}_s-\hat{P}_n)y}{y^T(I_n-\hat{P}_s)y} \cdot \frac{s-p}{n-s}$$

is the F-statistic (4.10) corresponding to (4.12). This has the central $F(n-s,\ s-p)$ distribution under the null hypothesis and is based upon the W-principle. Indeed (4.13) is obviously the Chow-test corresponding to the points $s = p+1,p+2,\ldots,n$. Writing $F_s$ for (4.13), it follows that

$$(4.14)\qquad Q_sF_s(n-s) = (1-Q_s)(n-s), \qquad s = p+1,p+2,\ldots,n-1.$$

Again, if for some $s$ $(n-s)>p$, it is more appropriate to apply the analysis of covariance rather than the prediction interval test, in which case we make the change corresponding to the one previously noted: in $\hat{P}_s$, replaced $I_{n-s}$ by the perpendicular projection on the span of the matrix comprising $(n-s)$ rows $x_t^T$ for $t = s+1,s+2,\ldots,n$.

Returning to the case $(n-s)\leqq p$, all of the above results arise from the fundamental orthogonal decomposition of the total sum of squares under the null hypothesis *i.e.* when $E(y_t) = x_t^T\beta = \mu_t$; writing $\mu$ for the $n \times 1$ vector of components $\mu_t$, the decomposition is:

$$(4.15)\qquad (y-\mu)^T(y-\mu) = (y-\mu)^T\hat{P}_n(y-\mu) + y^T(\hat{P}_{n-1}-\hat{P}_n)y + y^T(\hat{P}_{n-2}-\hat{P}_{n-1})y + \ldots$$

$$+ y^T(\hat{P}_{p+1}-\hat{P}_{p+2})y + y^T(I_n-\hat{P}_{p+1})y$$

which is readily re-written in terms of (4.9), namely

$$(4.16)\qquad y^T(I_n-\hat{P}_n)y = u_n^2 + u_{n-1}^2 + \ldots + u_{p+1}^2 \quad .$$

Thus both the cusum of squares test and the corresponding Chow-test are straightforward applications of the theory of the linear hypothesis, the former being based upon the M-principle and the latter upon the W-principle. Moreover, since there can be no conflict between these tests (provided the correct distributions are applied in each case), choice between them is entirely a matter of convenience. In this respect, the facility of easy sequential regression calculation that recursive residuals permit is clearly a great advantage. However, in applications of the TIMVAR program (Brown *et al.*, 1975), it must be remembered that the exact distribution of $Q_s$ is not used. There may thus be some conflict between the test results of TIMVAR and the corresponding calculations using the exact distributions of $Q_s$, or $F_s$.

In the above test all n observations are invariably used as the observations at t=s are varied over the range s = p+1,p+2,...,n. Harvey (1976) has considered a natural extension of the Chow-test when information is first available at time $s_1$, and then another $(s_2-s_1)$ observations become available; and then, after that, another $(s_3-s_2)$ bringing the total to $s_3$; and so on in blocks $s_i-s_{i-1}$, i = 1,2,...,q, say, where $s_q$=n. Harvey's test is a natural extension of the Chow-test in the sense it permits new observations to be tested but allows their number to vary at each stage. Moreover, the observations are naturally ordered in time and so, as we shall see, permit a decomposition of the total sum of squares into components which are familiar in the testing of several hypotheses by the method of 'pooling non-significant sums of squares' (see *e.g.* Seber, 1966, Chap. 6).

The general purpose is to test whether $x_t^T\beta_t$ has the form $\mu_t = x_t^T\beta$ for q blocks of observations: $s_1, s_2-s_1, \ldots, n-s_{q-1}$; $s_1 > p$. Corresponding to the blocks of observations, we may define perpendicular projections $\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_n$ corresponding to the matrices $\hat{X}_{s_1}, \hat{X}_{s_2}, \ldots, \hat{X}_n$, as defined in (4.4), it being presumed that for all i, $s_i - s_{i-1} \leqq p$. Consider the decomposition under $H_0$: $\beta_t = \beta$ for all t, corresponding to (4.15):

$$(4.17) \qquad (y-\mu)^T(y-\mu) = (y-\mu)^T\hat{P}_n(y-\mu) + y^T(\hat{P}_{q-1}-\hat{P}_n)y + y^T(\hat{P}_{q-2}-\hat{P}_{q-1})y + \ldots + y^T(\hat{P}_2-\hat{P}_3)y + y^T(\hat{P}_1-\hat{P}_2)y + y^T(I_n-\hat{P}_1)y \quad .$$

Based upon this decomposition, the following sequence of test statistics may be calculated:

$$(4.18) \qquad \frac{y^T(\hat{P}_1-\hat{P}_2)y}{y^T(I_n-\hat{P}_1)y}\cdot\frac{s_1-p}{s_2-s_1}, \quad \frac{y^T(\hat{P}_2-\hat{P}_3)y}{y^T(I_n-\hat{P}_2)y}\cdot\frac{s_2-p}{s_3-s_2} \quad , \ldots, \quad \frac{y^T(\hat{P}_{q-1}-\hat{P}_n)y}{y^T(I_n-\hat{P}_{q-1})y}\cdot\frac{s_{q-1}-p}{n-s_{q-1}} ,$$

all of which have F-distributions under the null hypothesis. Testing begins with the first factor in the sequence (4.18) and continues through the sequence until a significant value is obtained. Notice that the residual sum of squares in any denominator, save the first, is precisely the sum of the numerator and denominator of the previous ratio. For this reason, the procedure is seen to be one of 'pooling non-significant sums of squares'; and each test is essentially asking: "Given the previous $s_{i-1}$ observations belong to one regime, do the next $(s_i-s_{i-1})$ also belong to the same regime?" If each test in the sequence is not significant, a final check may be made by carrying out an F-test on the whole $(n-s_1)$ 'new' observations. So long as $(n-s_1) > p$, this F-test should be the analysis of covariance. Indeed, if at any point $s_i$, $(s_i-s_{i-1}) > p$, then the analysis of covariance may also be carried out at this stage. In this light, the decomposition (4.17) may always be cast in terms of perpendicular projections $\hat{P}_q$ which are either as given or are of a corresponding kind.

Let the blocks of $s_1$, $s_2-s_1$, $s_3-s_2$, ..., $s_q-s_{q-1}$ observations on $x_t^T$ be represented by $X_1$, $X_2$, $X_3$, ..., $X_q$ respectively. When the number of rows in any $X_i$ is greater than p, the corresponding perpendicular projection on its span is $P_i = X_i(X_i^TX_i)^{-1}X_i^T$; while if the number of rows is less than or equal to p, the corresponding perpendicular projection is taken to be the identity matrix of the same order as the number of rows. $\hat{P}_i$ is then block diagonal, comprising, in order, symmetric idempotent matrices, $P_i$ for some i and identity matrices as appropriate. This yields a series of perpendicular projections $\hat{P}_1$, $\hat{P}_2$, ... $\hat{P}_q$ $(\hat{P}_q=\hat{P}_n)$ and the decomposition (4.17) is again appropriate, as is the series of tests (4.18), after proper adjustment for degrees of freedom in the obvious way. By this device, the series of tests (4.18) becomes a mixture of the analysis of covariance and the prediction interval (or Chow-) test, and so represents a generalization of Harvey (1976).

## 5. *SUMMARY*

Three principles of testing hypotheses have been applied to the general linear hypothesis to clarify Seber's (1964) conclusion that the W-, M- and L-principles all "...lead to exactly the same test statistic". The W- and M-principles have been applied to tests for structural change in regression and it has been demonstrated that the Chow-test is based upon the W-principle while the cusum of squares is based upon the M-principle. Both tests are, consequently, straightforward applications of tests of a general linear hypothesis. For this reason, it has been possible to consider mixed tests and thereby to extend Harvey (1976). Moreover, while Harvey (1976) requires specialized proofs, the development here points to the underlying unity of several tests; by virtue of this, it avoids the need for specialized proofs and permits an easy extension of known results. Harvey's (1976) test has also been demonstrated to be a special case of the method of pooling non-significant sums of squares in the testing of several hypotheses.

## *6. ACKNOWLEDGEMENTS*

The author wishes to thank Michael McAleer for helpful discussions and comments on earlier drafts.

*REFERENCES*


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

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.

Fisher, G. (1978), "The Linear Hypothesis and Structural Change", *Working Paper*, Queen's University.

Fisher, G. (1980), "Propositions, Principles and Methods: The Case of the Linear Hypothesis", *Discussion Paper No. 377*, Institute for Economic Research, Queen's University.

Harvey, A.C. (1976), "An Alternative Proof and Generalization of a Test for Structural Change", *The American Statistician*, 30, 122-123.

Phillips, G.D.A. and A.C. Harvey (1974), "A Simple Test for Serial Correlation in Regression Analysis", *Journal of the American Statistical Association*, 69, 935-939.

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

Saw, J.G. (1964), "Some Notes on Variance-ratio Tests of the General Linear Hypothesis", *Biometrika*, 51, 511-518.

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

Seber, G.A.F. (1966), *The Linear Hypothesis: A General Theory*, London: Griffin.

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

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

Weatherburn, C.E. (1952), *A First Course in Mathematical Statistics*, Cambridge: C.U.P.

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