ISSN 0316-5078

INSTITUTE FOR ECONOMIC RESEARCH

QUEEN'S UNIVERSITY

Institute for economic research

Kingston, Ontario, Canada K7L 3N6

PROPOSITIONS, PRINCIPLES AND METHODS:

THE CASE OF THE LINEAR HYPOTHESIS

Gordon Fisher

Department of Economics
Queen's University

Discussion Paper No. 377

March, 1980

PROPOSITIONS, PRINCIPLES AND METHODS: THE CASE OF THE LINEAR HYPOTHESIS

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 then re-examined and put on a proper mathematical footing. Simple relations between the test statistics and their distributions are also outlined.

Gordon Fisher
March, 1980

PROPOSITIONS, PRINCIPLES AND METHODS: THE CASE OF THE LINEAR HYPOTHESIS

- Gordon Fisher -

1. *PREAMBLE*

The testing of propositions put forward hypothetically is fundamental to the advancement of knowledge in science, because it permits 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 entirely 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 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 potent guide to intuition and 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, namely, the Lagrange-multiplier test. 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, after appropriate adjustments, 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 ($p<n$). It is

desired to test H_0: $\theta \epsilon \omega$, where ω represents a sub-set of points in Ω which obey the r 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

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

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 $W \overset{a}{\sim} \chi^2(r)$. Similarly, if θ_Ω now refers to another asymptotic normal estimator of θ in Ω and $D_\Omega\{\cdot\}$ again denotes dispersion matrix corresponding to such estimation, 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 invariably based upon unrestricted estimation, that is, on 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 on estimation of θ in ω, that is, upon *restricted* estimation, using ϕ_ω, the estimate of the Lagrange multiplier corresponding to $h(\theta) = 0$. Of course, the large-sample test based upon the M-principle is given, in an obvious notation, by

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

which is again asymptotically $\chi^2(r)$ distributed under H_0. Notice carefully that there is no need to insist on maximum-likelihood estimation: the estimated Lagrange-multiplier ϕ_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. *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 leaves something to be desired. The discussion below attempts to give the conclusion a proper mathematical underpinning.

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:

$$(1) \qquad F = \frac{y^T(P_\Omega - P_\omega)y}{y^T(I_n - P_\Omega)y} \cdot \frac{n-p}{r}$$

where P denotes an orthogonal projection: P_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 x n matrix of rank $r \leq p$, then any $x \in \Omega$ which obeys $A^Tx = 0$ 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:

$$(2) \qquad P_\Omega - P_\omega = P_\Omega A(A^T P_\Omega A)^{-1} A^T P_\Omega \quad ,$$

provided $R[A] \cap \Omega^\perp$ comprises the origin only (Seber, 1964, p. 262). It is then easy to demonstrate that F in (1) embodies the W-principle since

$$(3) \qquad F = \frac{y^T(P_\Omega - P_\omega)y}{rs_\Omega^2} = \frac{(A^Tm_\Omega)^T[D_\Omega(A^Tm_\Omega)]^{-1}(A^Tm_\Omega)}{r}$$

where $D_\Omega(\cdot)$ denotes dispersion matrix evaluated at $\sigma^2 = s_\Omega^2$, the latter being given by $s_\Omega^2 = \{1/(n-p)\}\{y^T(I_n - P_\Omega)y\}$, and $m_\Omega = P_\Omega y$. Of course, F in (1) and (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_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 standardized normal variates, exactly under H_0. Since also s_Ω^2 is asymptotically equivalent to the maximum-likelihood estimator of σ^2, it is obvious that $rF \sim \chi^2(r)$ for large n. Corresponding to earlier notation, 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

$$(4) \qquad L = (y-\mu)^T(y-\mu) + 2\mu^T A\phi - 2\mu^T(I-P_\Omega)\kappa$$

for variations in μ and the vector Lagrange multipliers ϕ and κ in order that $(y-\mu)^T(y-\mu)$ is minimized, 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 (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 μ in ω. Writing f_ω for the estimate of ϕ corresponding to m_ω, the first-order conditions from (4) lead to:

$$(5) \qquad \begin{bmatrix} I_n & P_\Omega A \\ A^T P_\Omega & 0 \end{bmatrix} \begin{bmatrix} m_\omega \\ f_\omega \end{bmatrix} = \begin{bmatrix} P_\Omega y \\ 0 \end{bmatrix} .$$

This yields:

$$(6)\qquad \begin{bmatrix} m_\omega \\ f_\omega \end{bmatrix} = \begin{bmatrix} I_n - P_\Omega A(A^T P_\Omega A)^{-1} A^T P_\Omega & P_\Omega A(A^T P_\Omega A)^{-1} \\ (A^T P_\Omega A)^{-1} A^T P_\Omega & -(A^T P_\Omega A)^{-1} \end{bmatrix} \begin{bmatrix} P_\Omega y \\ 0 \end{bmatrix} ;$$

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

$$(7)\qquad s_\omega^2 = \frac{y^T(I_n - P_\omega)y}{n-p+r} = \frac{(y-m_\omega)^T(y-m_\omega)}{n-p+r}$$

is an unbiased estimate of σ^2 under H_0. The small-sample test statistic based upon the M-principle uses only estimates corresponding to estimation of μ in ω. If the statistic is M, then

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

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

The last expression is readily seen to reduce to

$$(8)\qquad M = \frac{y^T(P_\Omega - P_\omega)y}{y^T(I_n - P_\omega)y} \cdot (n-p+r)$$

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

$$(9)\qquad \frac{y^T(P_\Omega - P_\omega)y}{y^T(I_n - P_\omega)y} = \frac{y^T(P_\Omega - P_\omega)y}{y^T(P_\Omega - P_\omega)y + y^T(I_n - P_\Omega)y}$$

and the two components in the denominator of the right-hand side of (9), 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 (1) and (3), and M, in equation (8), may be written

$$(10) \qquad \frac{W}{n-p + W} = \frac{M}{n-p+r}$$

(see *e.g.* Weatherburn, 1952, chap. VIII; Wilks, 1962, p. 187). Moreover, if 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: $\lambda^{\frac{2}{n}} = \{\sigma^2_\Omega/\sigma^2_\omega\}$ where σ^2_Ω and σ^2_ω refer to maximum-likelihood estimates of σ^2 in Ω and ω, respectively. Thus, for large n,

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

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

$$(11) \qquad \lambda^{\frac{2}{n}} = \frac{M}{W} \cdot \frac{n-p}{n-p+r} \quad .$$

Note also that

$$(12) \qquad W = (\lambda^{-\frac{2}{n}} - 1)(n-p).$$

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 tests based upon the W-, M- and L-principles.

With regard to the calculated values of the test statistics, it is clear from (9), (10) and (12) that

$$(13) \qquad W = (\lambda^{-\frac{2}{n}} - 1)(n-p) \geqq M\{(n-p)/(n-p+r)\}$$

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

$$(14) \qquad W\{\frac{n}{n-p}\} \geqq \{-2 \log \lambda\} \geqq M\{\frac{n}{n-p+r}\} ,$$

each of which has the $\chi^2(r)$-distribution for large n. Relations (10) - (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 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 (1).

ACKNOWLEDGEMENT

I am most grateful to Michael McAleer for his encouragement and advice.

REFERENCES

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

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

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

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

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.

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

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