ASHLEY # A.1.10

A SIMPLE TEST FOR REGRESSION PARAMETER INSTABILITY

RICHARD ASHLEY*

The relative advantages of simple, exact tests over sophisticated but conservative tests are stressed and a simple diagnostic test to detect regression parameter instability is proposed. The test is exact (in the statistical sense), easy to perform using any regression package, and simple enough to present in a first year course. Thus, this test has the potential for widespread use. Monte Carlo simulation results are presented which indicate that the power of the test is comparable to that of more sophisticated alternatives — e.g., the VPR test of Garbade (1977). As an illustration, the test is applied to a bivariate forecasting model for Texas personal income constructed using time series analysis techniques.

I. INTRODUCTION

Much effort has been expended recently in the construction of sophisticated tests which detect regression parameter instability. The purpose of this paper is to propose a very simple test for parameter instability and to show that its performance is comparable to that of sophisticated alternatives.

Why is it useful to consider such a simple test? For one thing, the ordinary economist is unlikely to apply a diagnostic test which he has neither the background to understand nor the software to employ; more likely he will ignore the (possible) problem or test for it in an *ad hoc* manner. A simple but rigorous test which actually is applied is much more likely to correctly reject null hypotheses involving real data than a sophisticated test which is rarely or never applied. Moreover, the sophisticated tests often involve approximations; consequently, the actual size of these tests is often unknown, even asymptotically. This would seem to be a fatal embarrassment for a formal statistical test, except that the statistical community apparently is willing to accept tests for which only an upper bound on the size is known. Such tests are called "conservative."

Conservative tests can have high power in Monte Carlo simulations, where it is possible to adjust critical points to obtain, say, a 5% size. On real data, however, no such adjustment is possible. In that case a simple 5% test involving no approximations may systematically reject the null hypothesis more often than the conservative, sophisticated 5% test, merely because the conservative test has an actual size less than 5%. True, the probability of Type I error is smaller for the conservative test, but both results must, in all honesty, be reported and considered as 5% results. Consequently, it seems more sensible to compute and compare the nominal power of statistical tests, where nominal power is the probability of correctly rejecting the null hypothesis for a given nominal size. The nominal power of a test is, of course, identical to its power if and only if the test is exact — *i.e.*, if it has known size. Thus, a second reason for considering simple tests is that their nominal power can be comparable or superior to that of sophisticated, but approximate, tests.

*Economics Department, Virginia Polytechnic Institute and State University. The author would like to thank C. Spoede, C. C. Holt, J. A. Olson, and an anonymous referee for helpful comments.

II.

It was recognized quite early that one of the first casualties from the use of a faulty or incomplete economic model is the validity of the fixed coefficient assumption of the standard regression model, an assumption which is necessary in order to derive the optimality (efficiency, BLUness, *etc.*) of OLS regression. This recognition led to Quandt's (1958) work on regressions which switch from one set of fixed coefficients to another and to Chow's (1960) well-known test for a change in one or more coefficients between one part of the sample and the rest. [Maddala (1977, p. 198), lists still earlier references to the same test in the statistics literature. See also Farley and Hinich (1970) and Farley, Hinich, and McGuire (1975) for a simple alternative to the Chow test.]

More recently the profession has begun to consider alternatives to the fixed coefficient model where the parameters vary stochastically in every period. Cooley and Prescott (1973, 1976) introduced what they called the varying parameter regression model; there the k dimensional parameter vector, β_t, evolves in time according to a random walk:

$$(1.1) \qquad Y_t = X_t\beta_t + \epsilon_t \qquad t = 1, \ldots, N$$

$$(1.2) \qquad \epsilon_t \sim N(0, \sigma^2)$$

$$(1.3) \qquad \beta_t = \beta_{t-1} + u_t \qquad t = 1, \ldots, N$$

$$(1.4) \qquad u_t \sim N(0, \sigma^2 P),$$

where ϵ_t and u_t are independent white noise series and P is an exogenously given $(k \times k)$ matrix. Since u_t has mean zero, the parameters do not change on average; however, their variance grows linearly with time. An alternative model, which is perhaps more consonant with the existence of a meaningful economic theory underlying the specification, is Rosenberg's (1973) stochastically convergent parameter model, which appeared shortly thereafter. In Rosenberg's model

$$(1.5) \qquad \beta_t = (1-\lambda)\beta^* + \lambda\beta_{t-1} + u_t \qquad t = 1, \ldots, N,$$

where $0 < \lambda < 1$. Now β_t tends to converge to the fixed parameter vector β^* and ends up varying around β^* with a fixed steady state variance.

Stimulated by these developments, several new diagnostic tests for unstable regression coefficients have appeared recently, each designed to remain powerful in the context of stochastically varying parameters. The purpose of this paper is to introduce a simple test which has a number of desirable properties; first, however, each of these previously suggested tests is briefly reviewed.

The first test is due to Brown, Durbin and Evans (1975). They propose two tests, the "cusum" test and the "cusum of squares" test, both based on the recursive residuals $w_{k+1}, \ldots, w_N$. These residuals are the scaled one-step-ahead postsample forecast errors obtained by repeated ordinary least squares regression, increasing the sample size by one period at each repetition. Under the usual classical assumptions it can be shown that the w_t are independent normal variates with zero mean and unit variance.

The "cusum" test uses cumulative sums of the w_t series. Simulations in Garbade (1977) show that this test has little power to reject the null hypothesis of stable coefficients, so the cusum test will not be considered any further here. The "cusum of squares" test, which is based on cumulative sums of the squares of the w_t's is more powerful. It will be denoted "BDE" below.

The BDE test has three major drawbacks. First, it is neither exact, nor even conservative. For example, LaMotte and McWhorter (1980) find that a 5% BDE test wrongly rejects the null hypothesis of coefficient stability 7.3% of the time in their simulations. Second, the test inherently considers the stability of all k coefficients simultaneously; it is not possible to focus on a subset of the coefficients whose stability may be either in greater doubt or of more intense interest.[1] Finally, the results reported in section III below [and also similar results in LaMotte and McWhorter (1980) and Garbade (1977)] show that the power of the BDE test is substantially lower than that of the alternative tests discussed below. These results are plausible in view of Farley, Hinich and McGuire's (1975) proof that the BDE test is inconsistent.

The second recent diagnostic test is the varying-parameter regression test (VPR) introduced by Garbade (1977). He assumes the same coefficient variation model described in equations 1.1 to 1.4 above and tests the null hypothesis that the (kxk) matrix P is zero. This null hypothesis is equivalent to the assumption of stable coefficients.

When more than one coefficient is tested, the P matrix must be assumed to be a scalar multiple of some exogenously given matrix in order to proceed. Typically, this exogenous matrix is taken to be the identity matrix, but this choice seems quite arbitrary. It implies, for example, that the variance of each component of β_t grows linearly with time at the same rate. Under the null hypothesis, of course, it does not matter because this rate is zero.

The VPR test statistic does not have convenient statistical properties, however. For example, when k is one (so that P is a scalar) the likelihood ratio statistic is asymptotically distributed $\chi^2(1)$ under null hypotheses of the form: $H_0: P = P_0$ only when P_0 is greater than zero. For this reason, the actual size of the test is uncertain in practice even for large samples. The VPR test is conservative, in fact markedly so. For example, LaMotte and McWhorter (1980) find that a 5% VPR test wrongly rejects the null hypothesis only 1.0% of the time in their simulations.

So the VPR test is not particularly appealing on *a priori* grounds. On the other hand, the simulations in Garbade (1977) indicate that the VPR test is dramatically more powerful than both tests proposed by Brown, Durbin, and Evans (1975) in a variety of circumstances.

LaMotte and McWhorter (1978) have proposed an exact test which they designate LM. The LM test is based on the Cooley-Prescott random walk model of parameter variation also. As with the VPR test, the P matrix is expressed as a scalar multiple (σ_u^2) of an exogenously given matrix (D), usually taken (arbitrarily) to be an identity matrix. The null hypothesis that σ_u^2 is zero is then tested using a set of sufficient statistics invariant to changes in β_0, the starting value of the parameter vector.

LaMotte and McWhorter (1980) compare the power of the LM test to that of a

[1] The BDE test shares this disadvantage with the tests on the recursive residuals suggested by Harvey (1976).

number of likelihood ratio based alternatives, including one (FLR) which is equivalent to the VPR test, on a model with random walk parameter variation. They conclude that all of these tests have roughly the same power when adjusted to have the same actual size. (Recall that the actual size of a 5% VPR test is substantially less than 5% because it is not exact.) But such adjustments cannot be done in practice, so the unadjusted (nominal power) results in their table 1 are more relevant. Averaging over these results yields a power of 45% for the LM test versus 35% for the VPR test, so the LM test is more powerful in this instance than the VPR test. LaMotte and McWhorter prefer the LM test because it is also exact and involves no numerical optimization.

On the other hand, the LM test has drawbacks of its own. It does not have the intuitive appeal of the VPR test, for example. In addition, the LM test is invariant to changes in β_0 only for a given, fixed D matrix; from the discussion of the VPR test above, however, it should be clear that the appropriate D matrix is not invariant to scale changes in β_0. Suppose that the kth component of β_0 becomes ten times larger because the kth explanatory variable is now measured in different units. If D was a k dimensional identity matrix to begin with, then surely its kth diagonal element must now become one hundred; otherwise the degree of variation in the kth parameter will be distorted relative to the others. Thus, the LM test is fully invariant to changes in β_0 only in the special case where the test is separately applied to each coefficient by setting all of the elements of D to zero except for one diagonal element.

The LM test also requires a substantial amount of computation. First an orthonormal basis must be formed for the $N - k$ dimensional vector subspace orthogonal to the columns of the (Nxk) data matrix. Then the eigenvalues and eigenvectors of an $N - k$ dimensional matrix must be obtained. The expense of the test clearly grows quite quickly with the sample size, N. In addition, the VPR and BDE tests each provide a plot indicating the manner in which the coefficients appear to vary over the sample; the LM test does not.[2]

A simple test for unstable regression coefficients based on dummy variables is introduced below. It is

a. exact
b. easy to compute using only a standard regression package,
c. roughly as powerful (nominally) as the VPR test, and
d. so simple that it can easily be covered in a first year regression course.[3]

[2] A lagrange multiplier test has been proposed by Watson and Engle (1980) while this paper was in preparation. Their test has a more flexible alternative hypothesis than the LM and VPR tests, being based on Rosenberg's stochastically convergent markov model. On the other hand, the Watson/Engle test is quite conservative even for moderately large samples. [*E.g.*, they report (table 1) an average actual size for their 5% test of 1.08% for $N = 30$ and of 2.75% for $N = 100$.] So the nominal power of the test is probably comparable to that of the VPR and LM tests. In addition, the Watson/Engle test gives no picture of how the coefficients actually vary over the sample.

[3] This last point is somewhat double-edged. This test is *so* simple that it may be unfashionable to suggest it. Moreover, it must be noted explicitly that the author makes no claim as to having invented the use of dummy variables to detect parameter variation. The essential contribution of this paper is the assertion that the simple test proposed here compares favorably with the sophisticated tests in the recent literature.

In addition, it assumes little about the form of the instability, thus eliminating the need for separate tests against outliers, discrete parameter jumps, deterministic parameter drift, *etc.* This new test does yield a plot of the estimated parameter variation over time; the author calls this plot a "stabilogram," so the test itself is denoted "STAB" below. The STAB test is defined in section II. Its nominal power is investigated using Monte Carlo simulations in section III.

The stabilogram plays a role analogous to that of the correlogram in Box-Jenkins modeling — it can be used to test formally the need for further specification modifications and it can also be used informally to suggest the form such modifications should take. These uses are illustrated in section IV where stabilogram analysis is applied to a bivariate time series analysis model for forecasting the growth rate in personal income for Texas. Conclusions are noted in section V.

II. THE STABILOGRAM (STAB) TEST

The stabilogram test is a straightforward application of covariance analysis [*e.g.*, Johnston (1972), pp. 192-207]. Under the null hypothesis of stable coefficients the observations, y_t, are assumed to be generated by the standard linear model:

$$y_t = \sum_{\ell=1}^{k} \beta_\ell x_{\ell t} + u_t \qquad t = 1, \ldots, N. \tag{2.1}$$

$$u_t \sim N(0, \sigma^2), \tag{2.2}$$

where the $x_{\ell t}$ are exogenously given and the u_t are non-autocorrelated.

In its simplest form the test is applied to one coefficient at a time. There is no loss of generality in supposing that it is the kth coefficient whose stability is to be tested. The first step is to partition the sample period into r approximately equal subperiods of about $i = N/r$ observations each. Then r dummy variables, $D_t^{(1)}, \ldots, D_t^{(r)}$, are defined such that $D_t^{(1)}$ is one only in the first subperiod; $D^{(2)}$ is one only in the second subperiod, etc. Ordinary least squares regression is then applied to

$$y_t = \sum_{\ell=1}^{k-1} \beta_\ell x_{\ell t} + \sum_{j=1}^{r} \gamma_j D_t^{(j)} x_{tk} + u_t \qquad t = 1, \ldots, N. \tag{2.3}$$

The sequence of parameter estimates, $\hat{\gamma}_1, \ldots, \hat{\gamma}_r$, is the stabilogram of order i for the kth coefficient, or *STAB*(i, k).

Since each of these estimated coefficients is an estimator of β_k from a different subperiod, a picture of how β_k varies over time can be easily obtained by plotting a confidence interval around $\hat{\gamma}_j$ versus j. This is the "stabilogram" defined at the end of section I. This plot gives a visual indication as to what kind of parameter variation might be occurring. For example, where the confidence intervals trend upward substantially, one might tentatively conclude that a trending parameter is involved, and so forth.

It should-be noted that the stabilogram is not the only way to plot estimated parameter variation. Garbade's VPR test produces a plot of the estimated variation in the parameter also. In that plot the parameter estimates are based on an increasing number of observations as you move through the sample period. Brown, Durbin, and Evans (1975) have implemented a similar plot in their TIMVAR program; their "moving regressions" plot is the same as Garbade's except that they drop an early observation each time an additional observation is added, so that each parameter estimate in the plot is based on the same number of observations. Both of these plots are useful in that they appear to be more detailed than a stabilogram plot. On the other hand, they both require specialized software to produce efficiently and they both give a highly smoothed picture of the actual parameter variation. In contrast, the stabilogram smooths the variation only within each subperiod, so that the stabilogram will in general give a less highly smoothed picture of the parameter variation, with the degree of smoothing to a substantial degree under the user's control. The stabilogram also has the nice feature that the confidence intervals in it are exact, whereas the moving regressions plot does not give tolerance limits and those given by the VPR plot are only approximate.

The null hypothesis of stable coefficients corresponds to the $r - 1$ linear restrictions $\gamma_1 = \gamma_2 = \ldots = \gamma_r$. These may be easily tested using standard methods — *e.g.*, Maddala (1977, pp. 197-8). The appropriate test statistic is

$$(2.4) \qquad STAB = \frac{(RSS - URSS)/(r-1)}{URSS/(N-k-r+1)} ,$$

which is distributed $F(r-1, N-k-r+1)$ under the null hypothesis. *RSS* is the sum of squared residuals from equation 2.1 and *URSS* is the sum of squared residuals from equation 2.3.

This test is equivalent to the usual Chow test whenever there are just two subperiods. However, using such a small number of subperiods can be expected to yield a test of low power when the coefficients vary stochastically according to equation 1.3 or 1.5. On the other hand, there are insufficient degrees of freedom to compute $STAB\,(1, k)$.

Clearly the useful values of i lie somewhere in between these extremes. Greater resolution (smaller, more numerous subperiods) leads to fewer degrees of freedom in the estimated variance of u_t and consequently to fewer degrees of freedom in the denominator of the F statistic. In practice, subperiods of length five (*i.e.*, $i = 5$) seem to work well. But this choice is not critical — in the example described in section IV, subperiods of length twelve gave useful results.

The computational expense rises with r, the number of subperiods used, but only as fast as the cost of inverting the $r + k - 1$ dimensional data matrix $(X'X)$ used in OLS estimation of equation 2.3. Typically the value of r (and hence of i) is bounded by storage limitations in the regression package, not by computational expense. Note, however, that almost all of the computational burden and the storage requirements for large r can be eliminated if k is small by exploiting the simple (largely diagonal) structure of the $X'X$ matrix to obtain $(X'X)^{-1}$ analytically. Even when k is not small, the problem can be avoided by an appropriate partitioning of $X'X$ so that the largest non-diagonal matrix to be inverted is of order $k - 1$.

In some cases it is desirable or necessary to drop the restriction that the subperiods be of approximately equal length. For example, when one of the other $k - 1$ regressors is itself a dummy variable, a little caution must be exercised to avoid collinearity problems. And where the parameter variation (or one's interest in it) is localized in one part of the sample, that part can be examined at higher resolution by using smaller subperiods in that section. Note, however, that subperiods only one period long yield less interpretable results since the residual in that period will be forced to zero regardless of which coefficient is tested.

In still other cases one might want to test the stability of several coefficients at the same time. This, too, can be done (in an obvious way), at a further cost in resolution for any given sample size. Where the stability of all k coefficients is tested simultaneously, the *STAB* statistic is equivalent to the "homogeneity" statistic described in Brown, Durbin and Evans (1975, p. 156) and implemented in their TIMVAR program. Typically, however, the STAB test would be applied to only one or a few coefficients whose stability was of special interest. For example, it would not make much sense to examine the stability of the coefficient on a wartime dummy variable. Another reason for limiting attention to just one or a few coefficients is that multicollinearity in the explanatory variables may otherwise cause the confidence intervals plotted in the stabilogram plot to become uninterpretably large.

It must be noted, however, that the sequential application of the STAB test one coefficient at a time to more than one coefficient (in search of a significant result) is a form of data mining. As with all data mining, such a practice distorts the true significance level of the test. Consequently, it must be anticipated that some of the most suitable applications for the stabilogram test will be situations where either

a. there is enough data so that stabilogram dummy variables can be placed simultaneously on all coefficients of interest.

or

b. the stability of only a few coefficients is at issue or of interest.

For example, Spoede (1982) examines the stability of the parameter beta in the capital asset pricing model using stabilogram methods; an additional example is given below.

The STAB test is flexible, convenient, and simple; in addition it seems squarely focussed on the parameter variation issue. The critical question is, how powerful is the test against plausible forms of parameter instability? This question is taken up in the next section.

III. SIMULATION RESULTS

In this section the power of the stabilogram test, STAB *(i, k)*, defined above, is compared to that of the VPR, BDE, and Chow tests. The VPR and BDE tests are not exact, so the "power" of these tests as calculated and discussed in this section is understood to be "nominal power." *I.e.*, the tests are performed just as their originators described them. No effort is made to adjust artificially the critical points of these tests to make the actual size equal to the nominal size, since this kind of adjustment can never be made in practice with real data.

The comparison is made using Monte Carlo simulations designed to be directly comparable to those in which Garbade (1977) compared the power of the VPR and BDE tests. The model considered here is a simple one with just one explanatory variable:

$$(3.1) \qquad y_t = \beta_t x_t + \epsilon_t \qquad t = 1, \ldots, N$$

$$(3.2) \qquad \epsilon_t \sim N(0, 1)$$

$$(3.3) \qquad x_t \sim N(0, 5).$$

The x_t series was generated just once and held constant over all replications.[4] Following Garbade (1977), three patterns of stochastic coefficient instability were considered:

Case 1: *Random Walk*
(corresponds to equation 1.3 and to table 1)

$$\left(\begin{array}{ll} \beta_1 = 1.0 & \\ \beta_t = \beta_{t-1} + u_t & t = 2, \ldots, N \\ u_t \sim N(0, P) & \end{array}\right) \text{ for } \left(\begin{array}{c} P = .01, .10, 1.00 \\ \text{and} \\ N = 15, 31, 61 \end{array}\right)$$

Case 2: *Stable Markov Process*
(corresponds to equation 1.5 and to table 2)

$$\left(\begin{array}{ll} \beta_1 = 1.0 & \\ \beta_t = .7 + .3\beta_{t-1} + u_t & t = 2, \ldots, N \\ u_t \sim N(0, P) & \end{array}\right) \text{ for } \left(\begin{array}{c} P = .01, .10, 1.00 \\ \text{and} \\ N = 15, 31, 61 \end{array}\right)$$

Case 3: *Discrete Jump*

$$\left(\begin{array}{ll} \beta_t = 1 & 1 \leq t < .5(N-1) \\ \beta_t = 1 + (d/5N^{.5}) & .5(N-1) \leq t \leq N \end{array}\right) \text{ for } \left(\begin{array}{l} d = 1, 10, 100 \\ N = 15, 31, 61 \end{array}\right)$$

Simulation results are reported for each of these patterns in tables 1, 2, and 3, respectively. In each case the table gives the fraction of the trials in which a (nominal) 5% test rejected the null hypothesis of stable coefficients. The limiting factor on the number of replications was the expense of the numerical likelihood maximizations needed for the VPR test. This expense was greatly exacerbated by the frequent presence of multiple relative maxima. Consequently, the number of repetitions was set at two hundred.

[4] STAB simulations were repeated using a highly autocorrelated x_t series — AR(1) with $\phi = .80$ — the results were similar to those reported below.

TABLE 1
Random Walk Process
(Power on 5% Test)

P		N = 15	N = 31	N = 61
.01	STAB (2, 1)	33.0	60.5	95.0
	STAB (5, 1)	47.0	80.0	98.0
	VPR	38.0	63.0	90.0
	BDE	9.0	24.0	61.0
	Chow[a]	48.0	62.0	79.0
.10	STAB (2, 1)	79.5	97.0	100.0
	STAB (5, 1)	79.0	97.0	100.0
	VPR	82.0	98.5	100.0
	BDE	22.5	57.0	79.0
	Chow[a]	68.5	71.5	83.5
1.00	STAB (2, 1)	89.0	100.0	100.0
	STAB (5, 1)	86.0	100.0	100.0
	VPR	95.5	100.0	100.0
	BDE	35.5	62.0	82.5
	Chow[a]	78.0	78.0	87.5

[a] The Chow test splits the sample in half. It is thus equivalent to STAB (8, 1) for N = 15, to STAB (16, 1) for N = 31, and to STAB (3, 1) for N = 61.

TABLE 2
Stable Markov Process
(Power on 5% Test)

P		N = 15	N = 31	N = 61
.01	STAB (2, 1)	9.5	7.0	16.0
	STAB (5, 1)	7.5	9.5	20.5
	VPR	4.5	3.0	2.5
	BDE	9.5	7.0	9.5
	Chow[a]	6.5	6.5	6.5
.10	STAB (2, 1)	19.0	43.5	71.5
	STAB (5, 1)	23.5	49.5	78.0
	VPR	19.5	32.5	53.0
	BDE	14.0	15.5	31.5
	Chow[a]	15.5	18.0	26.5
1.00	STAB (2, 1)	29.5	71.0	94.5
	STAB (5, 1)	25.0	63.5	92.5
	VPR	34.5	72.5	94.0
	BDE	22.5	26.0	49.0
	Chow[a]	18.0	17.0	26.5

[a] See note "a" for Table 1.

TABLE 3

Discrete Jump
(Power on 5% Test)

d		N = 15	N = 31	N = 61
0	STAB (2, 1)	5.5	4.5	5.0
	STAB (5, 1)	3.0	6.0	6.5
	VPR	1.0	.5	0.0
	BDE	7.5	4.0	4.0
	Chow[a]	3.5	4.5	5.0
1	STAB (2, 1)	3.5	5.0	3.0
	STAB (5, 1)	5.5	9.0	4.5
	VPR	.5	0.0	0.0
	BDE	6.0	3.5	3.5
	Chow[a]	7.0	9.0	3.0
10	STAB (2, 1)	73.0	43.5	43.5
	STAB (5, 1)	98.0 [93.5][b]	87.0 [98.0][b]	76.5 [76.5][b]
	VPR	93.0 [91.0][b]	48.0 [73.0][b]	20.0 [18.0][b]
	BDE	13.0	20.0	18.5
	Chow[a]	100.0	95.5	97.5
	Garbade VPR[c]	65.0	79.7	88.0
100	STAB (2, 1)	100.0	100.0	100.0
	STAB (5, 1)	100.0	100.0	100.0
	VPR	100.0	100.0	100.0
	BDE	100.0	100.0	100.0
	Chow[a]	100.0	100.0	100.0

[a] See note "a" for Table 1.

[b] Figures in brackets are for an independent simulation using a different x_t series.

[c] These results are quoted from Garbade (1977); they are discussed in the text below.

If the true power of a particular test is p, then the observed power, $\hat{p}$, will be very nearly normally distributed with variance $p(1-p)/200$. The standard deviation of the sampling errors in tables 1, 2, and 3 thus varies from 1.5% (for $p = .05$) to 3.5% (for the worst case, $p = .50$). Since it seems likely that the sampling errors of the $\hat{p}$'s for the different tests are positively correlated, the standard errors of observed power differences in these tables are probably less than $(2)^{1/2}$ (1.5% - 3.5%), or 2% to 5%. Thus an observed power difference of 10% might be termed "significant at the 5% level" if both observed powers are close to .5, whereas an observed power difference of only 4% would suffice if both tests have an observed power close to one or zero.

An examination of the simulation results for random walk parameter variation (table 1) reveals that the BDE test is in every instance significantly less powerful than the other tests. The Chow test (a special case of the STAB test with just two subperiods) was the most powerful test only for the weakest parameter variation

($P = .01$) and smallest sample size; even then it is not significantly better than the STAB (5, 1) test. The STAB (5, 1) test appears to be significantly more powerful than the VPR test for weak parameter variation, but the VPR test looks somewhat better for strong parameter variation ($P = 1.0$). The STAB (2, 1) test does not appear to be an improvement over the STAB (5, 1) test for random walk parameter variation.

The simulation results for stable markov processes (table 2) indicate that both the BDE and the Chow tests are relatively weak; they perform comparably to the other tests only when the parameter variation is so weak ($P = .01$) and the sample size so small ($N = 15,31$) that none of the tests is very powerful. For weak parameter variation and $N = 61$ the STAB tests are significantly more powerful than any of the other tests. For moderate parameter variation ($P = .10$) the VPR and STAB tests are roughly equivalent at $N = 15$ and the STAB tests appear to be significantly superior for the larger sample sizes. As with the random walk case, the VPR test looks somewhat better than the STAB test for stronger parameter variation and the STAB (2, 1) test does not appear to be an improvement over the STAB (5, 1) test.

The simulation results for the discrete jump case are presented in table 3. Here none of the tests appear to be powerful for a very small jump ($d = 1$) and all of the tests yielded 100% for an extremely large jump.

For a moderate jump the Chow test is clearly superior, as one might expect since it is designed to detect this very type of parameter instability. [The Chow test would appear even more powerful, except that, as implemented here, it assumes that the jump takes place just after the midpoint of the sample (*i.e.*, period 8 when $N = 15$) whereas the actual jump takes place just before the midpoint (*e.g.*, period 7 when $N = 15$]. The BDE test is again weak relative to the other tests. The STAB (5, 1) test is not significantly less powerful than the Chow test for $N = 15$, but it is significantly less powerful for the larger sample sizes. The STAB (2, 1) test is distinctly less powerful than the STAB (5, 1) test in this case; clearly, this is an instance where the additional resolution provided by smaller, more numerous subperiods is superfluous and merely wastes degrees of freedom.

The VPR test appears to be rather weak in detecting discrete jump parameter variation for the larger sample sizes. This result is due to the discrepancy between the random walk parameter variation model on which the VPR test is based and the actual pattern present in this case. The observed power of the VPR test falls as the sample size increases for two reasons. For one thing, the size of the jump declines as N rises. Also, the larger samples have a larger number of pre-jump and post-jump observations over which the parameter really is stable. Garbade (1978) reported much more favorable results for the VPR test in the discrete jump case than were observed in the present study. These are presented in table 3 (for $d = 10$) with the label "Garbade VPR." His results are puzzling. It is not possible to replicate them because his x_t data are not available. However, it was possible to replicate the results presented here with an independently drawn set of x_t and ϵ_t data. These results are presented in brackets for the VPR and STAB (5, 1) tests in the $d = 10$ section of table 3. Notice that the change in the x_t's makes a substantial change in the figures for the $N = 31$ simulation,[5] but the relative pattern remains the same: the two tests are roughly equivalent for small samples, and both eventually decline in power as the

[5] Note that the true power of each test may depend on $x_1 \ldots x_t$, so it is not appropriate to compare these shifts to the standard errors of 2–5% discussed above.

sample size grows (and the jump shrinks), but the VPR test loses its power far more quickly than does the STAB (5, 1) test.

These simulation results indicate that the VPR test is slightly more powerful for detecting strong stochastic parameter variation and that the STAB tests are somewhat more powerful for detecting weak stochastic parameter variation. The STAB tests are superior for detecting discrete jumps.

While a high order STAB test *(e.g.*, the Chow test) is preferable for detecting a single discrete jump, its advantage over the STAB (5, 1) test would disappear if there were two jumps made during the sample period. The low order STAB tests [STAB (2, 1) and STAB (5, 1)] were clearly superior to the high order STAB tests in detecting stochastic parameter variation. Of the low order STAB tests, the STAB (5, 1) test seemed to work just as well as the STAB (2, 1) test on the stochastic variation cases and noticeably better on the discrete jump case, so the STAB (5, 1) test appears to be preferable. Overall, the STAB (5, 1) test appears to be the most powerful test considered here, with the VPR test a close second.

IV. AN ILLUSTRATIVE APPLICATION

In this section, stabilogram analysis is applied to a bivariate forecasting model for tpy_t, the quarterly growth rate in Texas personal income. Forecasts of tpy_t are of interest because a new constitutional amendment limits State appropriations for the 1982-83 biennium to γ times the expenditures in 1980-81, where γ is the accepted forecast of the ratio of average Texas personal income in the 1982-83 biennium to the same average over the 1980-81 biennium.

The forecasting model analyzed here was specified, estimated, and diagnostically checked using fairly standard time series analysis techniques based on the prewhitened cross-correlogram. [Details and data can be found in Ashley (1980)]. The general idea was to relate tpy_t to gnp_t, the growth rate in U.S. GNP, so that commercially available projections of gnp_t could be used to obtain forecasts of tpy_t through 1983.

The bivariate model was quite simple:

$$\begin{aligned} tpy_t = {} & \underset{(.52)}{.001636} + \underset{(1.43)}{.155}tpy_{t-1} + \underset{(.94)}{.104}tpy_{t-2} + \underset{(2.88)}{.285}tpy_{t-3} \\ & + \underset{(4.29)}{.483}gnp_t + e_t \qquad R^2 = .49 \\ & \qquad\qquad\qquad\qquad\quad DW = 2.02 \\ & \text{sample period} \quad 1961.2-1977.4 \end{aligned} \tag{4.1}$$

where the figures in parentheses are t statistics. This equation passed its in-sample diagnostic checks — *i.e.*, e_t appeared to be white noise uncorrelated with lagged, pre-whitened gnp_t. It also forecast satisfactorily over the post-sample period 1978.1 to 1979.4, reducing the root mean square error from .01290 (for a naive model based on a constant growth rate assumption) to .00641.[6]

[6] These forecasts were made using historical values for gnp_t; the 1978.1 forecast was a one-step-ahead forecast, the 1978.2 forecast was a two-step-ahead forecast, etc.

In short, equation 4.1 appeared to be so satisfactory that I invited two colleagues to apply the stabilogram test to it. [See C. C. Holt and J. A. Olson (1980) for details.] They chose to re-estimate equation 4.1 over the period 1960.1 to 1977.4 (obtaining similar results) and partitioned this sample period into six equal subperiods of twelve quarters each.

Their interest centered on the stability of the constant term and of the gnp_t coefficient. The stabilogram test results for these two coefficients are presented in table 4 below. In both cases the null hypothesis of parameter stability can be rejected at the .5% level.[7]

TABLE 4

Stabilogram Test Results on
Coefficients in the *tpy* Model
for Sample Period 1960.1–1977.4

Variable	coefficient	RSS	URSS	F(5,62)
constant	.420	.004285	.003080	4.85
gnp_t	.003	.004285	.003076	4.87

The stabilogram for the *gnp* coefficient is plotted in figure 1; this plot indicates that the *gnp* coefficient was substantially smaller in the 1960-62 subperiod (labelled "61" in figure 1) than it was subsequently. The stabilogram for the constant term shows a similar pattern. These stabilograms suggest that either

(a) the coefficients shifted in the early part of the sample period,

or

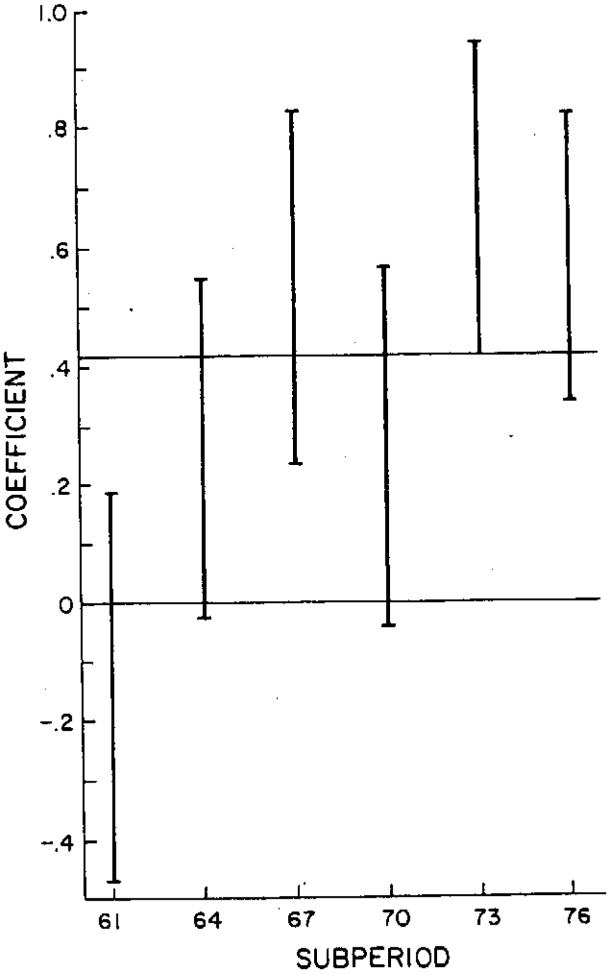
(b) the coefficients are drifting randomly.

The first of these two hypotheses seemes the most plausible one *a priori*, because the raw Texas personal income data from 1958 and 1959 had always looked a bit suspect to me. (In fact, the reason that the sample period for equation 4.1 was set to begin in 1961.2 was so that differenced models using dependent variables lagged up to four periods would never use any data prior to 1960.1.)

[7] Note that the stabilogram test is only asymptotically exact here due to the lagged dependent variables in equation 4.1. Also, the true significance level of the test is distorted by the fact that two separate stabilogram tests are done on the same regression.

FIGURE 1

Stabilogram on gnp_t Coefficient in tpy_t Model



Hypothesis (a) was tested by eliminating the first two subperiods (1960.1 to 1965.4) from the sample period. Stabilogram tests were then performed on both the constant term and on the gnp_t coefficient using the remaining four subperiods. The resulting test statistics, distributed F(3,40) under the null hypothesis, were 3.55 for the constant term and 5.64 for the gnp_t coefficient. Thus, the null hypothesis of stable coefficients is still rejected at the .5% level. Evidently, the removal of the early data enables the test to detect the more subtle parameter variation present in the remainder of the sample period.

Since eliminating the early data did not lead to a model with stable coefficients, it seems likely that the coefficients in this model are drifting randomly throughout the sample period. Holt and Olson (1980) have undertaken further analysis of equation 4.1 based on this assumption.

V. CONCLUSION

The simulations reported in section III above indicate that the stabilogram test (STAB) and the VPR test have approximately the same nominal power in detecting random walk and stable markov process parameter variation, with STAB having an edge for weak variation and VPR having an edge for strong variation. For discrete jump parameter variation the stabilogram test seems clearly superior to the VPR test except for very small samples where the two tests seem roughly equivalent in nominal power.

The simulations reported by LaMotte and McWhorter (1980) on a model with twenty observations and random walk parameter variation indicate a noticeably larger nominal power for the LM test than for the VPR test. It thus seems likely that the LM test is more powerful than the stabilogram test for detecting stochastic parameter variation. No doubt the stabilogram test is still the most powerful test for detecting discrete jump parameter variation.

On the other hand, the LM test requires quite lengthy computations and the LM test also provides no insight into the manner in which the coefficients vary, whereas the STAB test does through the stabilogram itself. The application to a model for Texas data in section IV demonstrates how essential this feature can be.

Moreover, even when the LM test does successfully detect parameter instability, the resulting conclusion — that the parameters are following a random walk — may be quite misleading. This feature is particularly significant since my experience with the stabilogram test suggests that economic data yields parameters that jump, shift, and/or trend as often as they take random walks.

In addition, the stabilogram test is inexpensive and easy to perform using just a standard regression package. It can be taught to everyone likely to need it as an illustrative example of linear hypothesis testing. If the effective power of a statistical test is defined as the nominal power multiplied by the probability that the test is actually applied, it seems clear that the simple test proposed here is clearly superior in terms of effective power to all tests that have been suggested so far.

REFERENCES

Ashley, Richard A., "Texas Forecasting Studies: Time Series Analysis," Bureau of Business Research, University of Texas at Austin, 1980, (BP 80-3).

Brown, R. L., Durbin, J., and Evans, J. M., "Techniques for Testing the Constancy of Regression Relationships Over Time, with Comments," *Journal of the Royal Statistical Society, Ser. B*, 1975, *37*, No. 2, 149-92.

Chow, Gregory, "Tests of Equality Between Subsets of Coefficients in Two Linear Regressions," *Econometrica*, *28*, 1960, 591-605.

Cooley, Thomas F., and Prescott, Edward C., "Estimation in the Presence of Stochastic Parameter Variation," *Econometrica*, 1976, *44*, 167-84.

____________, "An Adaptive Regression Model," *International Economic Review*, 1973, *14*, No. 2, June 1973, 364-371.

Farley, J. V., and Hinich, M., "Testing for a Shifting Slope Coefficient in a Linear Model," *Journal of the American Statistical Association* *65*, 1970, 1320-29.

Farley, J. V., Hinich, M., and McGuire, T. W., "Some Comparisons of Tests for a Shift in the Slopes of a Multivariate Linear Time Series Model," *Journal of Econometrics*, 1975, *3*, 297-318.

Garbade, Kenneth, "Two Methods for Examining the Stability of Regression Coefficients," *Journal of the American Statistical Society*, 1977, *72*, 54-63.

Harvey, A. C., "An Alternative Proof and Generalization of a Test for Structural Change," *The American Statistician*, 1976, *30*, 122-3.

Holt, Charles C., and Olson, Jerome A., "Texas Forecasting Studies: Structural Change," Bureau of Business Research, University of Texas at Austin, 1980, (BP 80-4).

Johnston, J., *Econometric Methods*, New York: McGraw-Hill, 1972.

LaMotte, Lynn R., and McWhorter, Archer, "An Exact Test for the Presence of Random-Walk Coefficients in a Linear Regression Model," *Journal of the American Statistical Society*, 1978, *73*, 816-820.

____________, "A Comparison of Tests for Random-Walk Regression Coefficients," 1980, unpublished manuscript.

Maddala, G. S., *Econometrics*, New York: McGraw-Hill, 1977.

Quandt, Richard, "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes," *Journal of the American Statistical Society*, 1958, *58*, 873-880.

Rosenberg, Barr, "The Analysis of a Cross Section of Time Series by Stochastically Convergent Parameter Regression," *Annals of Economic and Social Measurement*, 1973, *2*, 399-428.

Spoede, C. W., "An Investigation of Security Price Performance and Market Model Methodologies," 1982, unpublished manuscript.

Watson, Mark, and Engle, Robert F., "Testing for Varying Regression Coefficients When a Parameter is Unidentified Under the Null Hypothesis," University of California, San Diego Economics Department Discussion Paper, 1980.