GREDI
Groupe de Recherche en Économie
et Développement International

Cahier de Recherche / Working Paper
10-19

Kernel smoothing end of sample instability tests *P* values

Patrick Richard

UNIVERSITÉ DE SHERBROOKE

Kernel smoothing end of sample instability tests P values

Patrick Richard
GREDI and Département de sciences économiques
Faculté d'administration, Université de Sherbrooke
2500 Boulevard de l'Université
Sherbrooke, Québec, Canada
J1K 2R1
patrick.richard2@usherbrooke.ca
P:(819) 821-80000 ext. 63233
F:(819) 821-7934

Abstract

A Monte Carlo investigation shows that the rejection probability of the structural stability test of Andrews (2003) depends on several characteristics of the DGP, one of which is the length of the hypothesized break period. This is analyzed and found to be caused, at least in part, by the fact that the number of subsampling statistics used to compute the P value depends on the sample size and the length of the break period. Simulations show that kernel smoothed P values provide more accurate tests in small samples.

Keywords: Kernel smoothing; Simulation-based test; P value; Stability test.

JEL codes: C12; C14; C15.

This research was supported by a grant from the Fonds Québécois de Recherche sur la Société et la Culture.

March 2010

1 Introduction

Testing the null hypothesis of parameter stability is an important issue in econometric modeling. For example, suppose we posses $n+m$ observations of a variable y_t and of a set of k covariates X_t. Then, we may want to test the null hypothesis that the value of the elements of β is the same throughout the sample, that is, we may want to test

$$H_0 : \beta_0 = \beta_1$$

against

$$H_1 : \beta_0 \neq \beta_1$$

in the linear regression model

$$y_t = \begin{cases} X_t\beta_0 + u_t, & \text{for } t = 1, .., n \\ X_t\beta_1 + u_t, & \text{for } t = n+1, ..., n+m \end{cases}. \tag{1}$$

Several procedures have been proposed to perform such a test. Usually, their asymptotic characteristics under the null are derived under the assumptions that $n \to \infty$ and $m \to \infty$, that is, that both the pre and post break number of observations increases to infinity (see Andrews, 2003 and the references therein). This makes them somewhat inappropriate when the break period is thought to be short or when the break point occurs near the end of the sample.

Tests that only require $n \to \infty$ are developed by Dufour, Ghysels and Hall (1994) and Andrews (2003). Since these tests are closely related and because Andrews' (2003) critical values are much easier to obtain than those of Dufour et al. (1994), this paper is solely concerned with Andrews' (2003) test.

Essentially, Andrews' test is a variant of the well known F test proposed by Chow (1960). It is however, much more versatile since, in a linear regression when m is small and fixed, Chow's test necessitates normally distributed, independent and homoskedastic error terms to be asymptotically valid, whereas Andrew's test only requires stationarity and ergodicity. The S test statistic proposed by Andrews (2003) is defined as follows.

$$S = S_{n+1}(\hat{\beta}_{n+m}, \hat{\Sigma}_{n+m}), \tag{2}$$

where $\hat{\beta}_{n+m}$ is the OLS estimator of β computed using observations $t = 1, .., n+m$,

$$\hat{\Sigma}_{n+m} = \frac{1}{n+1} \sum_{j=1}^{n+1} \hat{u}_{j,j+m-1} \hat{u}_{j,j+m-1}^{\top}$$

$$\hat{u}_{j,j+m-1} = y_{j,j+m-1} - X_{j,j+m-1}\hat{\beta}_{n+m}$$

where $Z_{j,j+m-1}$ refers to rows $t = j, ..., j+m-1$ of the matrix Z. The function $S_{n+1}()$ is defined as

$$S_j(\beta, \Sigma) = A_j(\beta, \Sigma)^\top V_j^{-1}(\Sigma) A_j(\beta, \Sigma),$$

$$A_j(\beta, \Sigma) = X_{j,j+m-1}^\top \Sigma^{-1}(y_{j,j+m-1} - X_{j,j+m-1}\beta),$$

$$V_j(\Sigma) = X_{j,j+m-1}^\top \Sigma^{-1} X_{j,j+m-1}.$$

This statistic requires that the number of columns in the matrix X be smaller or equal to the number of post break observations ($k \leq m$). When this is not the case, Andrews proposes a second statistic, called P, defined as

$$P = P_{n+1}(\hat{\beta}_{n+m}, \hat{\Sigma}_{n+m}), \tag{3}$$

$$P_j(\beta, \Sigma) = (y_{j,j+m-1} - X_{j,j+m-1}\beta)^\top \Sigma^{-1}(y_{j,j+m-1} - X_{j,j+m-1}\beta).$$

When $k \geq m$, $S = P$. The P statistic can also be computed when $k \leq m$ but the simulations reported in Andrews (2003) indicate that the S test has better properties in these circumstances. Thus in what follows, I only consider the S test.

The P value of the S test is obtained by a method akin to subsampling. It is calculated using the empirical distribution function (EDF) of the statistics $S_j(\beta, \Sigma)$ for $j = 1, ..., n-m+1$. The logic of this method is evident: under both the null and the alternative, these statistics S_j are all calculated over subsamples that do respect the null hypothesis of structural stability. Thus, the EDF of S_j provides a good estimator of the distribution of S under the null. The S tests can easily be modified to carry out beginning and middle of sample instability tests. Exactly how to do this is explained in section 4 of Andrews (2003), to which the interested reader is referred.

2 Monte Carlo study

This section provides some Monte Carlo results on the S test's finite sample performances. A more detailed set of results can be found in Richard (2010). The matrix X is composed of a column of ones and $k = 1$, 2, 3 or 4 additional regressors generated from independent stationary AR(1) models with a parameter $\rho = 0,\ 0.4$

or 0.8. The error terms are also generated as an AR process with the same 3 possible values for ρ. Under the null hypothesis, I have set $\beta_0 = \beta_1 = 0$. The number of post break observations considered was 1, 5 or 10. Whereas Andrews (2003) only considers pre break sample sizes of 100 and 250, I used $n = 25$, 50, 100 and 250. The results presented in this section are based on 50 000 simulated samples.

The first column of table 1 shows the results of a response surface analysis wherein the S test's rejection probability (RP) was regressed on ρ, m and k. I ran different regressions for different sample sizes to clearly see how the effect of some of these factors changes as n increases. The effect of the sample size (n) may be indirectly seen through the constant of these regressions and is, obviously, to decrease the test's ERP. This was noted by Andrews (2003, p. 1685) and evidently results from standard asymptotic convergence arguments. Andrews (2003) also notes that larger values of ρ yield higher RPs. This feature appears in my simulations but the greater variety of sample sizes allows to clearly see that the magnitude of this effect is inversely related to n, as should be expected. The number of regressors, k, also has an impact on the test's RP which decreases as $n \to \infty$. Finally, the number of observations after the break, m, also affects the test's RP.

Part of the test's sensitivity to m may be due to the fact that the P value is calculated with $n - m + 1$ subsampling statistics. Indeed, in order for a simulation based test carried at a level α to be as precise as possible (and even exact if the statistic is a pivot), it is necessary that $\alpha(M + 1)$ be an integer, where M is the number of simulated statistics, see Davidson and MacKinnon (2000).

The curve labeled *S sub* in figure 1 illustrates this point. It shows the RP of the S test (evaluated with 1 000 000 simulated samples) as a function of n in a data generating process (DGP) with $k = 1$, $\rho = 0.4$ and $m = 5$. It is evident that the RP of the S test is very much influenced by the pair m and n. Notice that the troughs of the function correspond to samples sizes where $\alpha(n-m+2)$ is an integer ($n = 22, 42, 62, 82$ and, eventually, 102). What this means in practice is that the accuracy of the test does not monotonically increase with n and that adding an observation may cause the test to have a larger error in rejection probability (ERP). For instance, in the example presented in figure 1, the test has a RP of 0.0498 with $n = 42$ and 0.0653 when $n = 45$.

A solution to this problem is to compute kernel smoothed P values, that is, P values computed from a nonparametric estimate of the distribution of the S_j rather than from their EDF. This was first suggested by Racine and MacKinnon (2007) for computationnaly intensive bootstrap tests where computational cost may prevent

a large number of bootstrap replications to be used. Their simulations indicate that kernel smoothed bootstrap tests with noninteger $\alpha(M+1)$ have smaller ERPs than unsmoothed ones in small samples.[1]

An important element of this procedure is the choice of the bandwidth with which the smoothed P value is to be computed. Commonly used data-based rules for CDF estimation are $h_{IMSE} = 1.587sn^{-1/3}$ and $h_{MSE} = 1.3sn^{-1/3}$, where s is an estimate of the standard deviation of the subsampling tests statistics. These choices are, respectively, integrated mean squared error (IMSE) optimal and MSE-optimal. As noted by Racine and MacKinnon (2007), those selection techniques are not optimal for smoothed P value computations since they do not take into account the nominal level of the test performed. Using response surface analysis, they suggest $h_{R-MacK} = 1.575sB^{-4/9}$ for tests at 5% nominal level when the underlying distribution of the test is Gaussian.

Figure 1 plots the RP of the S test when the smoothed P value is computed using either one of these three methods. Obviously, the smoothed P values are much less affected by the sample size than the EDF-based ones. Furthermore, columns 2, 3 and 4 of table 1 provide response surfaces for each method. The dependence of the RP on m is weaker for the kernel smoothed P values. In addition, I have computed the root mean square error (RMSE) of the unsmoothed and smoothed tests RPs, that is, the RMSE of procedure i is defined as

$$RMSE_i = \frac{1}{J}\sum_{j=1}^{J}(RP_j^i - \alpha)^2,$$

where J is the total number of DGPs used. The smoothed P values provide tests with much smaller RMSE when $n = 25$ and comparable RMSE in all other sample sizes.

3 Conclusion

The simulation results presented here indicate that the length of the break period (m) is one of the factors that affect the RP of the S test proposed by Andrews (2003). At least part of this appears to be due to the fact that the number of subsampling statistics used to calculate the P value is a function of m. Kernel

[1] Their primary objective was to reduce the loss of power associated with simulation tests. Unreported simulations indicate that smoothed S tests have power similar to unsmoothed ones.

smoothing the distribution of the statistics before computing the P value successfully removes most of this dependence and yields more accurate RPs in small samples.

References

Andrews, D. W. K. (2003). "End-of-sample instability tests," *Econometrica*, **71**, 1661-1695.

Chow, G. C. (1960). "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, **28**, 591-605.

Davidson, R. and J. G. MacKinnon (2000). "Bootstrap tests: how many bootstrap?," *Econometric Reviews*, **19**, 55-68.

Dufour, J. M., E. Ghysels and A. Hall (1994). "Generalized predictive tests and structural change analysis in econometrics," *International Economic Review*, **35**, 199-229.

Racine, J. S. and J. G. MacKinnon (2007). "Inference via kernel smoothing of bootstrap P values," *Computational Statistics and Data Analysis*, **51**, 5949-5957.

Richard, P. (2010). "Accuracy of end of sample instability tests," Working paper, Université de Sherbrooke.

Table 1. Response surfaces

			$n = 25$	
	S sub	S IMSE	S MSE	S R-MacK
c	0.0822***	0.0609***	0.0636***	0.0690***
ρ	0.0302***	0.0252***	0.0261***	0.0280***
k	-0.0091***	-0.0080***	-0.0082***	-0.0080***
m	-0.0044***	-0.0028***	-0.0029***	-0.0033***
R^2	0.8075	0.8756	0.8731	0.8525
RP RMSE	0.0238	0.0170	0.0176	0.0192
			$n = 50$	
	S sub	S IMSE	S MSE	S R-MK
c	0.0585***	0.0547***	0.0559***	0.0570***
ρ	0.0253***	0.0234***	0.0238***	0.0249***
k	-0.0049***	-0.0046***	-0.0046***	-0.0043***
m	0.0000	-0.0009***	-0.0008***	-0.0006**
R^2	0.7496	0.7697	0.7649	0.7498
RP RMSE	0.0114	0.0111	0.0112	0.0112
			$n = 100$	
	S sub	S IMSE	S MSE	S R-MK
c	0.0470***	0.0500***	0.0506***	0.0513***
ρ	0.0173***	0.0173***	0.0174***	0.0177***
k	-0.0020***	-0.0022***	-0.0021***	-0.0020***
m	0.0004**	-0.0002	-0.0002	0.0000
R^2	0.7459	0.7208	0.7247	0.7305
RP RMSE	0.0072	0.0073	0.0073	0.0073
			$n = 250$	
	S sub	S IMSE	S MSE	S R-MK
c	0.0496***	0.0486***	0.0490***	0.0494***
ρ	0.0080***	0.0079***	0.0079***	0.0081***
k	-0.0005*	-0.0007**	-0.0006**	-0.0005
m	0.0004***	0.0001	0.0001	0.0002
R^2	0.7607	0.73	0.7284	0.7374
RP RMSE	0.0034	0.0032	0.0032	0.0032

, ** and * denote statistical significance at the 10%, 5% and 1% levels respectively according to heteroskedasticity-robust standard errors.*

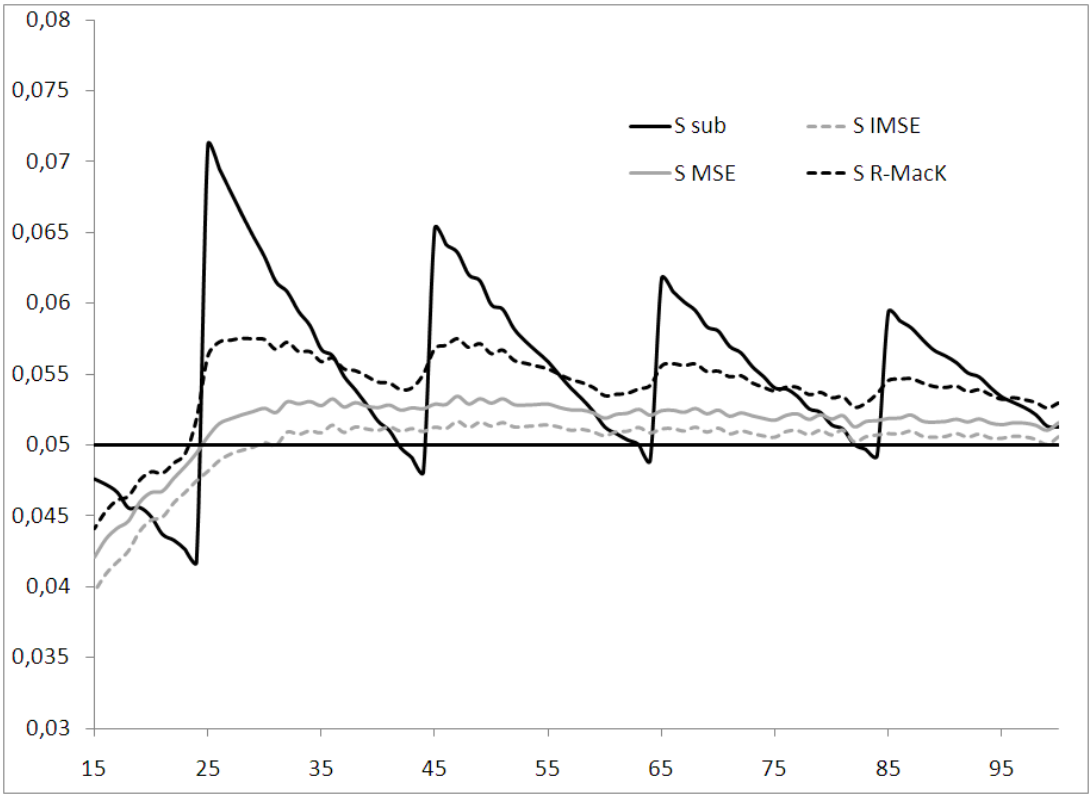

Figure 1: Rejection frequency of unsmoothed and smoothed S tests.