Economics Bulletin

Volume 30, Issue 1

The finite-sample properties of bootstrap tests in multiple structural change models

Jamel JOUINI
FSEGN and LEGI-EPT, Université 7 Novembre de Carthage, Tunisia

Mohamed Boutahar
Faculté des Sciences de Luminy, Université de la Méditerranée, France

Abstract

We propose bootstrap methods to approximate the distributions of test statistics for multiple structural breaks. The major advantage of these methods is that they allow freeing us from the constraints imposed by the asymptotic theory on parameters of the model. We also find that the asymptotic critical values lead to serious size distortions while, on the contrary, the bootstrap procedure leads to remarkably reliable tests in dynamic models.

Citation: Jamel JOUINI and Mohamed Boutahar, (2010) "The finite-sample properties of bootstrap tests in multiple structural change models", *Economics Bulletin*, Vol. 30 no.1 pp. 55-66.
Submitted: Jul 30 2009. **Published:** January 06, 2010.

1. Introduction

In the context of the use of the bootstrap to reduce the size distortions of tests, Diebold and Chen (1996) consider a simulation study to obtain better approximations to finite-sample distributions for single structural change tests using the bootstrap procedure. Their analysis presents numerous limitations since they don't allow for multiple breaks and shifts in the innovation variance across regimes. Here, we generalize their analysis so as to take these limitations into account. Indeed, we extend the analysis to multiple structural change tests[1] allowing for heterogeneity in the errors across regimes in the estimated regression model and propose some bootstrap tests as alternative solutions to the asymptotic tests constructed in Bai and Perron (1998).

The motivation for this paper lies in the fact that the asymptotic distribution theory of many of the break tests presented in the literature may not always be particularly useful in small sample situations. This is a practical difficulty because model selection problems such as the one of selecting the number of breaks in a time series is a small sample problem. The solution discussed in this paper is the one in which the asymptotic distribution is replaced by an empirical one, and the latter is obtained by a parametric or nonparametric bootstrap. We first show that the bootstrap methods allow yielding distributions independent of the minimal number of observations in each regime and the maximal possible number of breaks, unlike the asymptotic distributions which depend on these parameters as discussed by Bai and Perron (1998). We second show the accuracy of the bootstrap procedures and their ability to reduce or eliminate the error in rejection probability (size distortion) committed by the asymptotic tests.

The remainder of the paper is organized as follows. In section 2, we review the multiple structural change approach, namely the model, the estimation method and the test statistics. Section 3 introduces the bootstrap technique and the relating basic concepts being useful to carry out the simulation experiments. The finite-sample size performance of the asymptotic and bootstrap tests is investigated by simulations in section 4. We find that the bootstrap turns out to be a very useful tool in that it substantially improves size properties in dynamic models, regardless of the nature of the error distribution across regimes and whatever the value of the minimal number of observations in each segment. This contrasts with the results of Bai and Perron (2006) who find that the asymptotic tests are reliable only for large values of the trimming and homogenous errors across regimes. Section 5 concludes the paper.

2. The structural change approach

2.1 The model and the estimation method

The structural change model with m breaks, $(T_1, \ldots, T_m)$, is given by

[1] Considering multiple structural change tests is motivated by the fact that there is no theoretical justification for the assumption that a series would contain at the maximum one break and that there is no reason to suppose that the number of shifts will be known if the break locations are not. Moreover, it is well known that given the different international economic events a data series can be characterized by the presence of several breaks in its structure. Under these conditions, the tests designed to the case of only one break suffer from a power loss; and thus, considering multiple change tests becomes a need for exploring well the empirical evidence of instability in the series.

$$y_t = z_t'\delta_j + u_t, \tag{1}$$

where $t = T_{j-1}+1, \dots, T_j$ for $j = 1, \dots, m+1$, $T_0 = 0$ and $T_{m+1} = T$. y_t is the observed dependent variable, $z_t \in \mathbb{R}^q$ is the vector of regressors, δ_j is the corresponding vector of regression coefficients with $\delta_i \neq \delta_{i+1}$ $(1 \leq i \leq m)$, and u_t is the disturbance that has a distribution $D(0, \sigma^2)$, with $\sigma^2 < \infty$. The above model is expressed in matrix form as $Y = \bar{Z}\delta + U$, where $Y = (y_1, \dots, y_T)'$, $\delta = \left(\delta_1', \delta_2', \dots, \delta_{m+1}'\right)'$, $U = (u_1, \dots, u_T)'$, and $\bar{Z} = diag\,(Z_1, \dots, Z_{m+1})$ with $Z_i = \left(z_{T_{i-1}+1}, \dots, z_{T_i}\right)'$. For $i = 1, \dots, m$, we define the break fractions $\lambda_i = T_i/T$ with $0 < \lambda_1 < \cdots < \lambda_m < 1$. Bai and Perron (1998) impose some restrictions on the possible values of the break dates by defining the following set for some arbitrary small positive number ε: $\Lambda_\varepsilon = \{(\lambda_1, \dots, \lambda_m)\,; |\lambda_{i+1} - \lambda_i| \geq \varepsilon, \lambda_1 \geq \varepsilon, \lambda_m \leq 1 - \varepsilon\}$ to restrict each break date to be asymptotically distinct and bounded from the boundaries of the sample.

The estimation method proposed by Bai and Perron (1998) is based on the ordinary least-squares (OLS) principle. The method first consists in estimating the regression coefficients δ_j by minimizing the sum of squared residuals $\sum_{i=1}^{m+1} \sum_{t=T_{i-1}+1}^{T_i} (y_t - z_t'\delta_i)^2$. Once the estimate $\hat{\delta}\,(T_1, \dots, T_m)$ is obtained, we substitute it in the objective function and denote the resulting sum of squared residuals as $S_T\,(T_1, \dots, T_m)$. The estimated break dates $\left(\hat{T}_1, \dots, \hat{T}_m\right)$ are then determined by minimizing $S_T\,(T_1, \dots, T_m)$ over all partitions $(T_1, \dots, T_m)$ such that $T_i - T_{i-1} \geq [\varepsilon T]$, where $[\cdot]$ denotes integer part of argument. Finally, the estimated regression coefficients are such that $\hat{\delta} = \hat{\delta}\left(\hat{T}_1, \dots, \hat{T}_m\right)$. In the Monte Carlo Study, we use the efficient algorithm developed in Bai and Perron (2003a), based on the principle of dynamic programming, to estimate the unknown parameters.

2.2 The test statistics

We focus on tests for multiple breaks developed by Bai and Perron (1998) who first consider the sup F type test of structural stability against the alternative hypothesis that there is a known number of breaks k:

$$\sup F_T\,(k; q) = \sup_{(\lambda_1, \dots, \lambda_k) \in \Lambda_\varepsilon} F_T\,(\lambda_1, \dots, \lambda_k; q)\,, \tag{2}$$

where

$$F_T\,(\lambda_1, \dots, \lambda_k; q) = \frac{1}{T}\left(\frac{T - (k+1)\,q}{kq}\right)\hat{\delta}'R'\left(R\hat{V}\left(\hat{\delta}\right)R'\right)^{-1} R\hat{\delta}, \tag{3}$$

R is the conventional matrix such that $(R\delta)' = \left(\delta_1' - \delta_2', \dots, \delta_k' - \delta_{k+1}'\right)$, and $\hat{V}\left(\hat{\delta}\right)$ is an estimate of the variance covariance matrix of $\hat{\delta}$ given by $V\left(\hat{\delta}\right) = p\lim T\left(\bar{Z}'\bar{Z}\right)^{-1}\bar{Z}'\Omega\bar{Z}\left(\bar{Z}'\bar{Z}\right)^{-1}$, with $\Omega = E\,(UU')$.

The authors also consider double maximum tests of no structural change against an unknown number of breaks given some upper bound M for m:

$$D\max F_T\left(M,q,a_1,\ldots,a_M\right)=\max_{1\leq m\leq M} a_m \sup_{(\lambda_1,\ldots,\lambda_m)\in\Lambda_\varepsilon} F_T\left(\lambda_1,\ldots,\lambda_m;q\right). \tag{4}$$

Firstly, they set the weights $\{a_1,\ldots,a_M\}$ equal to unity, i.e. $a_m=1$, and label this version of the test as $UD\max F_T(M,q)$. Then, they consider a set of weights such that the marginal P values are equal across values of m. The weights are then defined as $a_1=1$ and for $m>1$ as $a_m=c(q,\alpha,1)/c(q,\alpha,m)$, where α is the significance level of the test and $c(q,\alpha,m)$ is the asymptotic critical value of the test $\sup F_T(m;q)$. This version of the test is denoted as $WD\max F_T(M,q)$.

Different versions of the tests can be obtained depending on the assumptions made with respect to the distribution of the regressors and the errors across subsamples. These relate to different specifications in the construction of the estimate of the matrix $V\left(\hat{\delta}\right)$ as discussed by Bai and Perron (2003a, 2006).

The asymptotic distributions of the above test statistics are derived in Bai and Perron (1998) and asymptotic critical values are tabulated in Bai and Perron (1998, 2003b) for $\varepsilon=0.05$ $(M=9)$, 0.10 $(M=8)$, 0.15 $(M=5)$, 0.20 $(M=3)$, and 0.25 $(M=2)$. For the present paper, the major advantage of the use of the bootstrap procedures is that they allow freeing us from the constraints imposed on the asymptotic distributions of the test statistics and their tabulations that depend on ε and M which, as we see, only take some particular values. Thus, to fill this gap we use the bootstrap distribution which allows affecting any value to these parameters.

3. Bootstrap procedure

Since the asymptotic tests may, in finite-samples, be biased in the sense that they have empirical sizes that differ from their nominal ones, an alternative approach is necessary. One such distribution-free method is the bootstrap method, introduced by Efron (1979). It is well known that the bootstrap procedure can successfully approximate the finite-sample distribution especially when the statistic is asymptotically pivotal as discussed previously (see Hall 1992, and Horowitz 1997). The considered test statistics are asymptotically pivotal since their asymptotic distributions do not depend on nuisance parameters as discussed by Bai and Perron (1998).

This paper proposes the following procedure to construct the bootstrapping distribution under the null hypothesis of structural stability;

1. Compute the asymptotic test statistic, say $\hat{\tau}$, in the usual way.[2]

2. Estimate the model under the null hypothesis by the OLS method and construct the bootstrap data-generating process (DGP) as follows:

$$y_t^*=z_t'\hat{\delta}+u_t^*, \qquad t=1,\ldots,T, \tag{5}$$

   where $\hat{\delta}$ is the restricted consistent OLS estimate of δ.

[2] $\hat{\tau}$ denotes the statistics $\sup F_T(k;q)$, $UD\max F_T(M,q)$ and $WD\max F_T(M,q)$.

- For the parametric bootstrap, we have $u_t^* \sim \hat{D}\left(0, \hat{\sigma}^2\right)$ where $\hat{\sigma}^2 = SSR/\left(T-q\right)$, with SSR is the sum of squared residuals. The parametric bootstrap uses a fully specified parametric model, which means that each set of parameter values defines just one DGP.
- The nonparametric bootstrap in the residuals of the fitted model is another way to perform the bootstrap in time series. The procedure is very similar to the parametric bootstrap, but now the distribution of the residuals is supposed to be unknown. In this situation, the errors u_t^* are generated by resampling with replacement from the residual vector $\tilde{u} = \{\tilde{u}_t\}_{t=1}^{T}$;

$$\tilde{u}_t = \sqrt{\frac{T}{T-q}} \left(\hat{u}_t - \frac{1}{T} \sum_{s=1}^{T} \hat{u}_s \right). \qquad (6)$$

3. Draw B bootstrap samples from the above DGP so as to obtain B bootstrap statistics, say $\left\{\hat{\tau}_j^*\right\}_{j=1}^{B}$, in exactly the same way as the real sample was used to compute the asymptotic statistic $\hat{\tau}$. Note that for the bootstrapping version of the test $WD\max F_T\left(M, q\right)$, the weights are $c^*\left(q, \alpha, 1\right)/c^*\left(q, \alpha, m\right)$, where $c^*\left(q, \alpha, m\right)$ is the bootstrap critical value obtained from the bootstrapping version of the test $\sup F_T\left(m; q\right)$.

4. Estimate the bootstrap P value, say $\hat{p}^*\left(\hat{\tau}\right)$, by the proportion of bootstrap samples that yield a statistic $\hat{\tau}^*$ greater than the asymptotic statistic $\hat{\tau}$. Analytically, let

$$\hat{G}^*\left(x\right) = P\left(\hat{\tau}^* \leq x\right) = B^{-1} \sum_{j=1}^{B} I\left(\hat{\tau}_j^* \leq x\right), \qquad (7)$$

denote the distribution function of the bootstrap test statistic $\hat{\tau}^*$, where $I\left(\hat{\tau}_j^* \leq x\right)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The bootstrap P value is then $\hat{p}^*\left(\hat{\tau}\right) = 1 - \hat{G}^*\left(\hat{\tau}\right)$.[3]

Unlike the asymptotic distributions, the bootstrap distribution $\hat{G}^*\left(x\right)$ depends on the potential specifications of the nature of the distributions for the errors and the regressors across subsamples since we can obtain various versions of the bootstrap tests as the asymptotic ones. Using this distribution, we can affect any value to the trimming ε and the maximum possible number of breaks M unlike the asymptotic distributions where these parameters only take some particular values (see subsection 2.2).

4. Monte Carlo analysis

4.1 Monte Carlo design

We report some Monte Carlo experiments to investigate the size of the tests for a Gaussian zero-mean first-order process without drift, which means that $z_t = y_{t-1}$ in the model (1). To that effect, we consider three ways of approximating the finite-sample distributions of

[3] In the case of fixed exogenous regressors and independent normal errors, the statistic $\hat{p}^*\left(\hat{\tau}\right)$ can be interpreted as an exact Monte Carlo P value as discussed previously (see Dwass 1957).

the statistics; the first approximation is Bai and Perron's (1998) asymptotic distribution, the second approximation is the parametric bootstrap distribution and the last one is the nonparametric bootstrap distribution.

Note that the bootstrap samples must be constructed recursively because of the presence of a lagged dependent variable. This is necessary because y_t^* must depend on y_{t-1}^* and not on y_{t-1} from the observed data. The recursive rule for generating a bootstrap sample is

$$y_t^* = \hat{\delta} y_{t-1}^* + u_t^*, \qquad t = 1, \ldots, T. \tag{8}$$

Note that for the parametric bootstrap $u_t^* \sim NID\left(0, \hat{\sigma}^2\right)$, where $\hat{\sigma}^2$ is a consistent estimator of σ^2 whereas for the nonparametric one, the errors u_t^* are generated by resampling with replacement from the residuals given by (6).

For each procedure, we attempt to see how the probabilities of rejecting the null hypothesis vary with the persistence as measured by the coefficient on the lagged dependent variable, the trimming ε and the distribution of the errors across regimes in the estimated regression model for a sample size $T = 150$. The trimming and the maximum possible number of breaks are as follows: $\varepsilon = 0.05$ $(M = 5)$, 0.10 $(M = 5)$, 0.15 $(M = 5)$ and 0.20 $(M = 3)$.[4] The nominal level of the tests is $\alpha = 5\%$, the number of Monte Carlo replications is set at $N = 500$ and we choose $B = 199$.[5] This choice of B is justified by the fact that we are interested in doing inference under the null hypothesis. It follows from Hall (1986) that the error in rejection probability made by a bootstrap test is $O\left(T^{-(j+1)/2}\right)$ (for some integer $j \geq 1$), regardless of the number of bootstrap samples used to estimate the bootstrap critical values.

4.2 Results of the asymptotic approximation

The results are reported in Table 1. The examination of the finite-sample properties of the asymptotic approximation illustrates that the size distortions are severe and huge since the tests tend to over-reject when the serial correlation increases and we allow for heterogeneity in the errors across regimes. The tendency of the tests to over-reject decreases as the trimming increases.[6] When the degree of persistence is small and for all values of the trimming, the asymptotic tests slightly under-reject when there is no heterogeneity in the errors across subsamples since the actual test size tends to be pushed downward relative to nominal size. When $\delta \leq 0.5$, the distribution of the errors quite greatly affects the empirical size of the tests, whereas when serial correlation is high, this distribution has a slight effect on the size performance of the tests especially for large values of ε.

Other simulation experiments are carried out for $T = 200$, 250.[7] The obtained results indicate that the size distortions of the tests are still severe even often for some selected parameter values for which the asymptotic tests may not encounter problems. Indeed, for $T = 250$, $\varepsilon = 0.20$ and $\delta = 0.75$, the maximum difference between the actual and nominal

[4] Note that other values for the trimming ε and the maximum permitted number of breaks M are possible if our sole concern is the bootstrap tests.

[5] The value of B is chosen to satisfy the condition that $\alpha(B + 1)$ is an integer as discussed by Davidson and MacKinnon (2000). This choice of B deletes all eventual bias of the bootstrap estimation of a P value.

[6] Note that allowing for different variances of the errors induces substantial size distortions unless the trimming is large for high values of δ.

[7] The results are not reported, but are available upon request from the authors.

sizes is 7.8% for the 5%-level test when we do not allow for heterogeneity in the errors across regimes. A nearly unit-root greatly affects the empirical size even for large samples.

Since the finite-sample performance of the asymptotic tests is not good, it is judicious to find another approximation to the finite-sample distribution so as to correct the deficiencies of the asymptotic distribution. This leads us to the bootstrap methods.

4.3 Results of the bootstrap procedures

The results are presented in Tables 2 and 3. The bootstrap techniques consistently outperform the asymptotic distribution in approximating the finite-sample distribution since they perfectly reduce the differences between the actual and nominal sizes even for some chosen parameter values for which the bootstrap tests may encounter problems. Indeed, with the nonparametric bootstrap approximation and $\varepsilon = 0.05$, the maximum difference between the empirical and nominal sizes is 1.9% for the 5%-level test when the errors are homogenous across regimes. Thus, for our examined model, the bootstrap eliminates the problem of excessive finite-sample size of the tests. The bootstrap performance is nearly perfect even when we allow for heterogeneity in the errors across regimes regardless of the value of the nuisance parameter δ with slight superiority to the nonparametric bootstrap. Thus, we are now able to choose large values of the persistence degree to do correct inferences. Consequently, there is a marked contrast between the performance of asymptotic tests and their bootstrap counterparts. Indeed, the use of tests based on the asymptotic approximation is associated with severe size distortions which heavily depend on δ.

The results show that the bootstrap methods improve on asymptotic approximations and quasi-perfectly solve the inference problem since the error in rejection probability is very minimal even when we allow for shifts in the innovation variance across subsamples. Thus, the bootstrap distribution is a good approximation to the finite-sample distribution even when the parametric serial correlation is high and the distribution of the errors differs across regimes.

4.4 Comments

- If heterogeneity in the errors across regimes is or not allowed in the estimated regression model, the size of the asymptotic tests is adequate under the hypothesis of no break for large sample sizes and trimming, and small autoregressive coefficient.
- In most cases, the asymptotic tests don't achieve the right size for small parametric serial correlation, so there is no reason to be surprised that they fail to provide good size properties for near-nonstationary cases.
- The bootstrap tests are accurate and often much more accurate than the asymptotic tests since they have a very small error in rejection probability. This is the reason for using bootstrap tests instead of asymptotic tests.
- The fact that the bootstrap procedures are more efficient than the asymptotic approximations is confirmed by examining the differences between the empirical sizes and the nominal level of the tests. Indeed, the bootstrap presents on the whole values of order 10^{-3}, while for the inference based on the asymptotic distributions the difference is about 10^{-1} in most cases.

- We have seen that bootstrapping has reduced the size distortions of the tests. This is because we know that if an asymptotic test over-rejects under the null hypothesis, a bootstrap test based on it will reject less often both under the null and under many alternatives as discussed by MacKinnon (2002).

- Conversely, if an asymptotic test under-rejects under the null hypothesis, a bootstrap test based on it will reject more often. This can be shown by the results corresponding to the cases where $\delta = 0.05$ for all values of ε, and there is no heterogeneity in the errors across subsamples.

- We have carried out some Monte Carlo experiments in which we have allowed for different distributions of the regressors across subsamples. The results (not reported but available upon request from the authors) indicate that for the case of no heterogeneity in the errors, the asymptotic tests have high tendency to under-reject under the null hypothesis. However, for the case of heterogeneity, they have tendency to over-reject but less than for the case where the regressors are homogenous across regimes. The bootstrap tests show the same size properties as the case of homogeneity in the regressors across regimes.

- In general, it is known that bootstrap tests seem to perform extremely well in the context of models with exogenous regressors and *i.i.d.* errors. We observe here that in the context of testing for multiple structural changes, bootstrap tests also perform well for dynamic models.

5. Conclusion

This paper has presented bootstrap methods and illustrated their ability to overcome the problem of incorrect finite-sample size of multiple structural change tests for dynamic models. The results of the tests are affected by the change of some factors; and the nonparametric bootstrap appears to perform slightly better than the parametric one for structural change tests applied to AR(1) process. The results are of interest since they provide evidence on the accuracy and adequacy of bootstrap methods to eliminate the problem of size distortions committed by the tests when the inference is based on the asymptotic methods.

References

[1] Bai, J and P. Perron (1998) "Estimating and testing linear models with multiple structural changes" *Econometrica* **66**, 47-78.

[2] Bai, J and P. Perron (2003a) "Computation and analysis of multiple structural change models" *Journal of Applied Econometrics* **18**, 1-22.

[3] Bai, J and P. Perron (2003b) "Critical values for multiple structural change tests" *Econometrics Journal* **1**, 1-7.

[4] Bai, J and P. Perron (2006) "Multiple structural change models: a simulation analysis" in *Econometric Theory and Practice: Frontiers of Analysis and Applied Research* by D. Corbea, S. Durlauf and B.E. Hansen, Eds., Cambridge University Press, 212-237.

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

[6] Diebold, F.X and C. Chen (1996) "Testing structural stability with endogenous break point: a size comparison of analytic and bootstrap procedures" *Journal of Econometrics* **70**, 221-241.

[7] Dwass, M. (1957) "Modified randomization tests for nonparametric hypotheses" *Annals of Mathematical Statistics* **28**, 181-187.

[8] Efron, B. (1979) "Bootstrap methods; another look at the Jackknife" *Annals of Statistics* **7**, 1-26.

[9] Hall, P. (1986) "On the number of bootstrap simulations required to construct a confidence interval" *Annals of Statistics* **14**, 1453-1462.

[10] Hall, P. (1992) *The Bootstrap and Edgeworth Expansion*, Springer Verlag: New York.

[11] Horowitz, J.L. (1997) "Bootstrap methods in econometrics: theory and numerical performance" in *Advances in Economics and Econometrics: Theory and Applications* by D.M. Kreps and K.F. Wallis, Eds.,Cambridge University Press, 188-222.

[12] MacKinnon, J.G. (2002) "Bootstrap inference in econometrics" *Canadian Journal of Economics* **35**, 615-645.

Table 1. Empirical size of the asymptotic tests

$T = 150$		No heterogeneity in the errors					Heterogeneity in the errors				
		δ					δ				
ε	Tests	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
0.05	$\sup F_T\,(1)$	4.4	5.6	9.0	18.0	36.8	10.0	11.0	13.6	21.4	38.8
	$\sup F_T\,(2)$	2.4	4.0	9.8	31.4	62.8	12.2	15.0	24.8	39.2	65.2
	$\sup F_T\,(3)$	1.2	3.0	9.0	30.2	67.4	14.2	15.6	25.6	42.2	73.6
	$\sup F_T\,(4)$	2.8	3.4	7.8	33.4	73.6	14.2	17.8	27.2	54.2	79.2
	$\sup F_T\,(5)$	1.8	3.2	7.6	34.6	80.6	16.2	17.8	28.2	55.4	84.8
	$UD\max$	3.4	6.2	12.0	37.0	80.0	18.0	19.6	31.8	54.8	85.8
	$WD\max$	3.6	5.8	12.4	42.6	84.0	21.6	22.2	38.2	60.6	90.8
0.10	$\sup F_T\,(1)$	4.6	5.4	8.0	12.6	32.0	6.6	7.6	9.8	14.2	33.2
	$\sup F_T\,(2)$	2.0	3.4	8.8	23.0	55.6	6.8	7.2	16.6	26.0	59.8
	$\sup F_T\,(3)$	2.4	3.6	6.8	23.6	56.2	8.2	9.6	14.8	28.2	58.8
	$\sup F_T\,(4)$	2.0	2.8	6.2	23.4	61.0	6.8	9.4	15.8	30.4	66.0
	$\sup F_T\,(5)$	1.8	2.2	6.2	21.6	63.6	7.4	8.8	17.0	30.6	68.6
	$UD\max$	4.0	6.0	10.2	23.8	64.2	7.8	10.6	18.6	31.0	68.6
	$WD\max$	2.8	5.2	9.6	28.6	70.2	9.0	11.4	21.6	37.8	75.4
0.15	$\sup F_T\,(1)$	4.8	5.4	7.0	10.8	27.2	5.8	6.4	8.2	11.2	27.8
	$\sup F_T\,(2)$	2.2	4.2	6.4	18.0	44.8	6.2	5.2	10.6	20.6	47.6
	$\sup F_T\,(3)$	2.4	3.2	6.4	16.4	45.4	5.4	7.0	10.2	19.0	49.4
	$\sup F_T\,(4)$	2.2	2.8	5.5	16.4	47.0	6.0	6.6	10.2	21.6	51.0
	$\sup F_T\,(5)$	2.4	2.6	5.6	17.2	49.2	7.0	7.4	10.2	22.8	52.6
	$UD\max$	4.2	5.0	8.2	16.2	51.0	7.2	7.2	11.0	21.0	53.2
	$WD\max$	4.0	4.8	8.4	20.6	55.4	8.2	8.2	13.8	24.8	59.8
0.20	$\sup F_T\,(1)$	3.6	4.2	6.2	9.0	22.2	4.2	5.6	7.6	10.2	23.6
	$\sup F_T\,(2)$	2.2	3.2	6.2	13.6	35.2	4.2	4.0	7.6	14.6	36.8
	$\sup F_T\,(3)$	2.4	3.2	5.4	12.4	34.4	4.0	4.8	7.0	14.4	37.6
	$UD\max$	3.4	4.4	6.8	13.6	36.8	5.8	6.6	8.8	16.0	40.0
	$WD\max$	3.6	4.2	6.8	14.8	41.0	6.0	6.4	9.2	16.6	42.2

Table 2. Empirical size of the parametric bootstrap tests

$T = 150$		No heterogeneity in the errors					Heterogeneity in the errors				
		δ					δ				
ε	Tests	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
0.05	$\sup F_T\,(1)$	3.7	3.9	4.0	4.2	5.0	3.0	4.0	3.9	4.6	4.6
	$\sup F_T\,(2)$	4.6	3.9	3.9	3.8	3.0	2.9	3.9	4.8	4.0	3.8
	$\sup F_T\,(3)$	4.2	3.9	4.4	5.4	3.9	3.8	4.2	4.4	4.4	4.4
	$\sup F_T\,(4)$	4.5	4.0	4.8	3.8	4.0	3.8	4.6	3.8	3.4	5.2
	$\sup F_T\,(5)$	4.8	4.2	5.0	4.2	4.1	4.7	5.0	4.4	5.4	3.9
	$UD\max$	4.5	4.6	4.4	4.4	4.2	3.9	5.0	4.3	4.6	3.8
	$WD\max$	5.8	5.6	6.0	5.0	4.1	5.8	7.4	6.4	6.2	5.4
0.10	$\sup F_T\,(1)$	4.1	3.6	5.2	5.2	5.8	6.0	4.7	5.8	5.2	6.4
	$\sup F_T\,(2)$	6.2	5.4	5.5	5.4	6.2	6.4	7.4	7.0	6.4	5.0
	$\sup F_T\,(3)$	6.2	7.0	6.0	5.0	4.4	6.6	8.0	8.1	5.2	7.0
	$\sup F_T\,(4)$	6.2	6.6	6.0	4.8	4.6	7.2	8.0	7.1	4.9	6.6
	$\sup F_T\,(5)$	7.0	6.4	6.1	4.5	4.8	6.8	7.3	7.8	5.6	6.6
	$UD\max$	4.8	5.0	5.0	4.8	6.4	7.0	7.1	6.2	5.8	6.8
	$WD\max$	5.8	5.6	5.2	4.8	6.2	9.8	10.1	10.0	8.4	7.4
0.15	$\sup F_T\,(1)$	4.6	4.2	6.2	4.8	6.6	4.2	5.0	7.0	4.4	6.0
	$\sup F_T\,(2)$	4.8	4.6	5.6	5.4	6.6	6.2	6.0	7.0	4.8	5.8
	$\sup F_T\,(3)$	5.8	5.0	5.4	5.8	6.8	6.2	5.6	7.4	4.6	5.4
	$\sup F_T\,(4)$	5.4	5.0	5.0	5.2	5.2	6.0	5.1	7.4	4.8	5.2
	$\sup F_T\,(5)$	5.4	4.8	5.0	5.8	4.8	5.8	5.4	6.8	4.6	5.0
	$UD\max$	4.2	4.0	5.6	5.6	6.4	6.0	4.8	6.6	4.1	6.2
	$WD\max$	4.0	3.9	6.0	5.4	6.6	8.2	5.2	6.8	4.0	9.0
0.20	$\sup F_T\,(1)$	4.6	4.8	4.8	5.4	6.0	4.6	4.6	5.4	4.8	5.8
	$\sup F_T\,(2)$	3.8	3.6	5.8	5.4	8.0	5.0	7.2	9.0	6.0	7.2
	$\sup F_T\,(3)$	4.2	4.0	5.4	5.8	7.2	4.6	4.6	5.4	5.3	6.6
	$UD\max$	4.0	5.0	5.0	5.4	6.4	4.4	4.6	6.2	5.2	7.2
	$WD\max$	3.9	3.9	5.4	6.0	6.6	5.0	5.0	7.0	5.6	7.6

Table 3. Empirical size of the nonparametric bootstrap tests

$T = 150$		No heterogeneity in the errors					Heterogeneity in the errors				
		δ					δ				
ε	Tests	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
0.05	$\sup F_T\,(1)$	5.2	4.7	4.7	3.9	4.5	4.6	4.2	4.8	4.2	4.6
	$\sup F_T\,(2)$	6.0	6.5	5.0	5.1	5.2	6.3	6.4	5.2	5.2	4.5
	$\sup F_T\,(3)$	5.4	5.2	4.6	5.0	5.2	5.6	5.9	5.6	4.6	4.6
	$\sup F_T\,(4)$	4.9	5.4	6.2	5.0	4.9	6.0	5.8	6.8	6.0	4.2
	$\sup F_T\,(5)$	4.6	4.9	5.0	4.6	5.8	5.6	5.1	6.2	4.9	5.6
	$UD\max$	5.6	5.2	5.0	5.2	5.6	5.2	4.6	7.4	5.0	5.2
	$WD\max$	6.9	6.4	6.4	5.1	6.0	8.0	8.6	10.4	7.4	7.2
0.10	$\sup F_T\,(1)$	3.6	4.0	4.4	4.0	5.6	3.4	4.0	4.0	3.6	5.6
	$\sup F_T\,(2)$	4.4	5.4	5.6	4.8	5.6	6.0	4.8	5.8	7.4	5.4
	$\sup F_T\,(3)$	4.4	6.1	5.2	6.0	5.0	5.0	5.8	6.0	6.4	5.8
	$\sup F_T\,(4)$	4.6	5.0	5.6	4.4	4.6	4.2	5.4	4.8	5.6	5.0
	$\sup F_T\,(5)$	4.6	5.2	4.9	4.6	5.4	4.6	4.6	4.2	5.6	6.2
	$UD\max$	4.8	4.8	3.9	4.4	5.2	4.0	5.1	5.6	5.4	5.2
	$WD\max$	5.2	4.6	3.9	3.8	5.8	6.0	7.0	7.1	6.5	7.0
0.15	$\sup F_T\,(1)$	4.2	3.8	3.9	5.0	7.2	3.8	4.4	3.9	4.8	7.0
	$\sup F_T\,(2)$	4.4	5.2	6.0	4.6	5.0	4.4	5.6	5.8	5.6	5.6
	$\sup F_T\,(3)$	4.8	6.0	6.1	5.8	5.0	5.8	6.0	6.0	6.0	5.0
	$\sup F_T\,(4)$	4.2	4.6	5.4	4.8	4.6	5.0	5.2	5.6	6.0	6.0
	$\sup F_T\,(5)$	4.0	4.6	4.2	4.5	5.0	4.2	5.0	5.4	5.4	6.4
	$UD\max$	4.6	4.6	4.5	4.4	6.0	4.8	4.6	5.2	5.5	5.8
	$WD\max$	4.0	4.6	4.9	4.6	5.8	7.2	6.6	6.2	8.0	7.2
0.20	$\sup F_T\,(1)$	4.7	3.6	4.4	4.6	8.2	3.4	3.4	4.2	4.8	8.2
	$\sup F_T\,(2)$	4.6	4.4	5.2	5.2	6.2	4.2	4.0	4.6	4.9	6.2
	$\sup F_T\,(3)$	3.9	5.4	5.4	5.0	6.6	4.6	4.8	5.2	4.4	6.2
	$UD\max$	4.4	4.4	5.0	4.6	7.0	4.2	3.8	4.2	4.6	6.1
	$WD\max$	3.9	4.4	4.6	5.8	6.1	4.8	5.4	5.0	6.2	8.0