Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Journal of Econometrics 122 (2004) 317 – 347
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

Simulation-based finite-sample tests for heteroskedasticity and ARCH effects

Jean-Marie Dufour[a,*], Lynda Khalaf[b,c], Jean-Thomas Bernard[c], Ian Genest[c]

[a] *CIRANO, CIREQ, and Département de Sciences Économiques, Université de Montréal, Montréal Qué., Canada H3C 3J7*
[b] *CIREQ, Univeristé de Montréal, Montréal, Qué., Canada H3C 3J7*
[c] *GREEN, Université Laval, Pavillon J.-A.-De Sève, St. Foy, Qué., Canada G1K 7P4*

Accepted 9 October 2003

Abstract

Tests for heteroskedasticity in linear regressions are typically based on asymptotic approximations. We show that the size of such tests can be perfectly controlled in finite samples through Monte Carlo test techniques, with both Gaussian and non-Gaussian (heavy-tailed) disturbance distributions. The procedures studied include standard heteroskedasticity tests [e.g., Glejser, Bartlett, Cochran, Hartley, Breusch–Pagan–Godfrey, White, Szroeter] as well as tests for ARCH-type heteroskedasticity. Sup-type and combined tests are also proposed to allow for unknown breakpoints in the variance. The fact that the proposed procedures achieve size control and have good power is demonstrated in a Monte Carlo simulation.

JEL classification: C12; C15; C2; C52; C22

Keywords: Linear regression; Specification test; Monte Carlo test; Exact test; GARCH

1. Introduction

Detecting and making adjustments for the presence of heteroskedasticity in the disturbances of statistical models is one of the fundamental problems of econometric methodology. We study the problem of testing the homoskedasticity of linear

* Corresponding author. Départment de Sciences Économiques, C.P. 6128 succursale Centre-ville, Montréal, Québec, Canada H3C 3J7. Tel.: +1-514-343-2400; fax: +1-514-343-5831.

E-mail address: jean.marie.dufour@umontreal.ca (J.-M. Dufour).

URL: http://www.fas.umontreal.ca/SCECO/Dufour

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2003.10.024

regression disturbances, under parametric (possibly non-Gaussian) distributional assumptions, against a wide range of alternatives, especially in view of obtaining more reliable or more powerful procedures. The heteroskedastic schemes we consider include random volatility models, such as ARCH and GARCH error structures, variances which are functions of exogenous variables, as well as discrete breaks at (possibly unknown) points.

The statistical and econometric literatures on testing for heteroskedasticity are quite extensive.[1] In linear regression contexts, the most popular procedures include the Goldfeld–Quandt F-test (Goldfeld and Quandt, 1965), Glejser's regression-type tests (Glejser, 1969), Ramsey's versions of the Bartlett (1937) test (Ramsey, 1969), the Breusch–Pagan–Godfrey Lagrange multiplier (LM) test (Godfrey, 1978; Breusch and Pagan, 1979), White's general test (White, 1980), Koenker's studentized test (Koenker, 1981), and Cochran–Hartley-type tests against grouped heteroskedasticity (Cochran, 1941; Hartley, 1950; Rivest, 1986); see the literature survey results in Table 1.[2]

The above methods do not usually take variances as nuisance parameters that must be taken into account (and eventually eliminated) when making inference on other model parameters (such as regression coefficients). More recently, in time series contexts and especially financial data analysis, the modeling of variances (volatilities) as a stochastic process has come to be viewed also as an important aspect of data analysis, leading to the current popularity of ARCH, GARCH and other similar models.[3] As a result, detecting the presence of conditional stochastic heteroskedasticity has become an important issue, and a number of tests against the presence of such effects have

Table 1
Survey of empirical literature on the use of heteroskedasticity tests

Heteroskedasticity test used	Literature share (%)
Tests for ARCH and GARCH effects	25.3
Breusch–Pagan–Godfrey–Koenker	20.9
White's test	11.3
Goldfeld–Quandt	6.6
Glejser's test	2.9
Hartley's test	0.3
Other tests	1.9
Use of heteroskedasticity consistent standard errors	30.3

Note: This survey is based on 379 papers published in The Journal of Business and Economic Statistics, The Journal of Applied Econometrics, Applied Economics, the Canadian Journal of Economics, Economics Letters, over the period 1980–1997. These results were generously provided by Judith Giles.

[1] For reviews, the reader may consult Godfrey (1988), Pagan and Pak (1993) and Davidson and MacKinnon (1993, Chapters 11 and 16).

[2] Other proposed methods include likelihood (LR) tests against specific alternatives [see, for example, Harvey (1976), Buse (1984), Maekawa (1988) or Binkley (1992)] and "robust procedures", such as the Goldfeld and Quandt (1965) peak test and the procedures suggested by Bickel (1978), Koenker and Bassett (1982) and Newey and Powell (1987).

[3] See Engle (1982, 1995), Engle et al. (1985), Bollerslev et al. (1994), LeRoy (1996), Palm (1996), and Gouriéroux (1997).

been proposed; see Engle (1982), Lee and King (1993), Bera and Ra (1995) and Hong and Shehadeh (1999).

Despite the large spectrum of tests available, the vast majority of the proposed procedures are based on large-sample approximations, even when it is assumed that the disturbances are independent and identically distributed (i.i.d.) with a normal distribution under the null hypothesis. So, there has been a number of recent studies that seek to improve the finite-sample reliability of commonly used homoskedasticity tests. In particular, Honda (1988) and Cribari-Neto and Ferrari (1995) derived Edgeworth and Bartlett modifications for the Breusch–Pagan–Godfrey criteria, while Cribari-Neto and Zarkos (1999) proposed bootstrap versions of the latter procedures. Tests based on the jackknife method have also been considered; see Giaccotto and Sharma (1988) and Sharma and Giaccotto (1991).[4]

A limited number of provably exact heteroskedasticity tests, for which the level can be controlled for any given sample size, have been suggested. These include: (1) the familiar Goldfeld–Quandt F-test and its extensions based on BLUS (Theil, 1971) and recursive residuals (Harvey and Phillips, 1974), which are built against a very specific (two-regime) alternative; (2) a number of procedures in the class introduced by Szroeter (1978), which also include Goldfeld–Quandt-type tests as a special case (see Harrison and McCabe, 1979; Harrison, 1980, 1981, 1982; King, 1981; Evans and King, 1985a); (3) the procedures proposed by Evans and King (1985b) and McCabe (1986). All these tests are specifically designed to apply under the assumption that regression disturbances are independent and identically distributed (i.i.d.) according to a normal distribution under the null hypothesis. Further, except for the Goldfeld–Quandt procedure, these tests require techniques for computing the distributions of general quadratic forms in normal variables such as the Imhof (1961) method, and they are seldom used (see Table 1).

Several studies compare various heteroskedasticity tests from the reliability and power view-points.[5] In addition, most of the references cited above include Monte Carlo evidence on the relative performance of various tests. The main findings that emerge from these studies are the following: (i) no single test has the greatest power against all alternatives; (ii) tests based on OLS residuals perform best; (iii) the actual level of asymptotically justified tests is often quite far from the nominal level: some are over-sized (see, for example, Honda, 1988; Ali and Giaccotto, 1984; Binkley, 1992), while others are heavily under-sized, leading to important power losses (see Lee and King, 1993; Evans, 1992; Honda, 1988, Griffiths and Surekha, 1986; Binkley, 1992); (iv) the incidence of inconclusiveness is high among the bounds tests; (v) the exact tests compare favorably with asymptotic tests but can be quite difficult to implement

[4] In a multi-equations framework, Bewley and Theil (1987) suggested a simulation-based test for a particular testing problem; however, they did not supply a distributional theory, either exact or asymptotic.

[5] See, for example, Ali and Giaccotto (1984), Buse (1984), MacKinnon and White (1985), Griffiths and Surekha (1986), Farebrother (1987), Evans (1992), Godfrey (1996), and, in connection with GARCH tests, Engle et al. (1985), Lee and King (1993), Sullivan and Giles (1995), Bera and Ra (1995) and Lumsdaine (1995).

in practice. Of course, these conclusions may be influenced by the special assumptions and simulation designs that were considered.

In this paper, we describe a general solution to the problem of controlling the size of homoskedasticity tests in linear regressions. We exploit the technique of Monte Carlo (MC) tests (Dwass, 1957; Barnard, 1963; Jöckel, 1986; Dufour and Kiviet, 1996, 1998) to obtain provably exact randomized analogues of the tests considered. This simulation-based procedure yields an exact test whenever the distribution of the test statistic does not depend on unknown nuisance parameters (i.e., it is *pivotal*) under the null hypothesis. The fact that the relevant analytical distributions are quite complicated is not a problem: all we need is the possibility of simulating the relevant test statistic under the null hypothesis. In particular, this covers many cases where the finite-sample distribution of the test statistic is intractable or involves parameters which are unidentified under the null hypothesis, as occurs in the problems studied by Davies (1977, 1987), Andrews and Ploberger (1995), Hansen (1996) and Andrews (2001). Further the method allows one to consider any error distribution (Gaussian or non-Gaussian) that can be simulated.

This paper makes five main contributions to the theory of regression based homoskedasticity tests. First, we show that all the standard homoskedasticity test statistics considered [including a large class of residual-based tests studied from an asymptotic viewpoint by Pagan and Hall (1983)] are pivotal in finite samples, hence allowing the construction of finite-sample MC versions of these.[6] In this way, the size of many popular asymptotic procedures, such as the Breusch–Pagan–Godfrey, White, Glejser, Bartlett, and Cochran–Hartley-type tests, can be perfectly controlled for any parametric error distribution (Gaussian or non-Gaussian) specified up to an unknown scale parameter.

Second, we extend the tests for which a finite-sample theory has been supplied for Gaussian distributions, such as the Goldfeld–Quandt and various Szroeter-type tests, to allow for non-Gaussian distributions. In this context, we show that various bounds procedures that were proposed to deal with intractable finite-sample distributions (e.g., by Szroeter, 1978; King, 1981; McCabe, 1986) can be avoided altogether in this way.

Third, our results cover the important problem of testing for ARCH, GARCH and ARCH-M effects. In this case, MC tests provide finite-sample homoskedasticity tests against standard ARCH-type alternatives where the noise that drives the ARCH process is i.i.d. Gaussian, and allow one to deal in a similar way with non-Gaussian disturbances. In non-standard test problems, such as the ARCH-M case, we observe that the MC procedure circumvents the unidentified nuisance parameter problem.

Fourth, due to the convenience of MC test methods, we define a number of new test statistics and show how they can be implemented. These include: (1) combined Breusch–Pagan–Godfrey tests against a break in the variance at an unknown point; (2) combined Goldfeld–Quandt tests against a variance break at an unspecified point, based

[6] For the case of the Breusch–Pagan test, the fact that the test statistic follows a null distribution free of nuisance parameters has been pointed out by Breusch and Pagan (1979) and Pagan and Pak (1993), although no proof is provided by them. The results given here provide a rigorous justification and considerably extend this important observation.

on the minimum (*sup*-type) or the product of individual p-values; (3) extensions of the classic Cochran (1941) and Hartley (1950) tests, against grouped heteroskedasticity, to the regression framework using pooled regression residuals. Although the null distributions of many of these tests may be quite difficult to establish in finite samples and *even asymptotically*, we show that the tests can easily be implemented as finite-sample MC tests. [7]

Fifth, we reconsider the notion of "robustness to estimation effects" (see Godfrey, 1996, section 2) to assess the validity of residual-based homoskedasticity tests. In principle, a test is considered robust to estimation effects if the underlying asymptotic distribution is the same irrespective of whether disturbances or residuals are used to construct the test statistic. Our approach to residual-based tests departs from this asymptotic framework. Indeed, since the test criteria considered are pivotal under the null hypothesis, our proposed MC tests will achieve size control for any sample size, even with non-normal errors, whenever the error distribution is specified up to an unknown scale parameter. Therefore, the adjustments proposed by Godfrey (1996) or Koenker (1981) are not necessary for controlling size.

The paper makes several further contributions relevant to empirical work. Indeed, we conduct simulation experiments [modelled after several studies cited above including: Honda, 1988; Binkley, 1992; Godfrey, 1996; Bera and Ra, 1995; Lumsdaine, 1995] which suggest new guidelines for practitioners. Our results first indicate that the MC versions of the popular tests typically have superior size and power properties, which motivates their use particularly in ARCH or break-in-variance contexts. Second, whereas practitioners seem to favor Breusch–Pagan–Godfrey type tests, Szroeter-type tests clearly emerge as a better choice (in terms of power). In the same vein, our proposed variants of Hartley's test—although the latter test is not popular in econometric applications—appear preferable to the standard LR-type tests (in terms of power versus application ease). We also provide guidelines regarding the number of MC replications.

The paper is organized as follows. Section 2 sets the statistical framework and Section 3 defines the test criteria considered. In Section 4, we present finite sample distributional results and describe the Monte Carlo test procedure. In Section 5, we report the results of the Monte Carlo experiments and with Section 6 we conclude.

2. Framework

We consider the linear model

$$y_t = x_t'\beta + u_t, \tag{1}$$

$$u_t = \sigma_t \varepsilon_t, t = 1, \ldots, T, \tag{2}$$

where $x_t = (x_{t1}, x_{t2}, \ldots, x_{tk})', X \equiv [x_1, \ldots, x_T]'$ is a full-column rank $T \times k$ matrix, $\beta = (\beta_1, \ldots, \beta_k)'$ is a $k \times 1$ vector of unknown coefficients, $\sigma_1, \ldots, \sigma_T$ are (possibly random) scale parameters, and

[7] For example, the combined test procedures proposed here provide solutions to a number of *change-point* problems. For further discussion of the related distributional issues, the reader may consult Shaban (1980), Andrews (1993) and Hansen (2000).

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)' \text{ is a random vector with a completely specified continuous distribution conditional on } X. \tag{3}$$

Clearly, the case where the disturbances are normally distributed is included as a special case. We are concerned with the problem of testing the null hypothesis

$$\mathrm{H}_0\colon \sigma_t^2 = \sigma^2, t = 1, \ldots, T, \text{ for some } \sigma, \tag{4}$$

against the alternative $\mathrm{H}_A\colon \sigma_t^2 \neq \sigma_s^2$, for at least one value of t and s. More precisely, we consider the problem of testing the hypothesis that the observations were generated by a data-generating process (DGP) which satisfies assumptions (1)–(4).

The hypothesis defined by (1)–(4) does not preclude dependence nor heterogeneity among the components of ε. No further regularity assumptions are assumed, including the existence of the moments of $\varepsilon_1, \ldots, \varepsilon_T$. So we can consider heavy-tailed distributions such as stable or Cauchy, in which case it makes more sense to view tests of H_0 as tests of scale homogeneity. So in most cases of practical interest, one would further restrict the distribution of ε, for example by assuming that the elements of ε are independent and identically distributed (i.i.d.), i.e.

$$\varepsilon_1, \ldots, \varepsilon_T \text{ are i.i.d. according to some given distribution } F_0, \tag{5}$$

which entails that $u_1, \ldots, u_T$ are i.i.d. with distribution function $\mathsf{P}[u_t \leqslant v] = F_0(v/\sigma)$ under H_0. In particular, it is quite common to assume that

$$\varepsilon_1, \ldots, \varepsilon_T \overset{\text{i.i.d}}{\sim} \mathrm{N}[0, 1], \tag{6}$$

which entails that $u_1, \ldots, u_T$ are i.i.d. $\mathrm{N}[0, \sigma^2]$ under H_0. However, as shown in Section 4, the normality assumption is not needed for several of our results; in particular, it is not at all required for the validity of MC tests for general hypotheses of the form (1)–(4), hence, *a fortiori*, if (4) is replaced by the stronger assumption (5) or (6).

We shall focus on the following special cases of heteroskedasticity (H_A), namely:

H_1 : GARCH and ARCH-M alternatives;
H_2 : σ_t^2 increases monotonically with one exogenous variable $(x_1, \ldots, x_T)'$;
H_3 : σ_t^2 increases monotonically with $E(y_t)$;
H_4 : σ_t^2 is the same within p subsets of the data but differs across the subsets; the latter specification is frequently termed *grouped heteroskedasticity*.

Note that H_4 may include the hypothesis that the variance changes discretely at some (specified) point in time. We also propose exact tests for a structural break in the variance at unknown points. In most cases, the tests considered are ordinary least squares (OLS) based. For further reference, let:

$$\hat{\sigma}^2 = \hat{u}'\hat{u}/T, \ \hat{u} = (\hat{u}_1, \ldots, \hat{u}_T)' = y - X\hat{\beta}, \quad \hat{\beta} = (X'X)^{-1}X'y. \tag{7}$$

3. Test statistics

The tests we shall study, which include existing and new procedures, can be conveniently classified in three (not mutually exclusive) categories: (i) the general class of tests based on an auxiliary regression involving OLS residuals and some vector of explanatory variables z_t for the error variance; (ii) tests against ARCH-type alternatives; (iii) tests against grouped heteroskedasticity. Unless stated otherwise, we shall assume in this section that (6) holds, even though the asymptotic distributional theory for several of the proposed procedures can be obtained under weaker assumptions.

3.1. Tests based on auxiliary regressions

3.1.1. Standard auxiliary regression tests

To introduce these tests in their simplest form (see Pagan and Hall, 1983), consider the following auxiliary regressions:

$$\hat{u}_t^2 = z_t'\alpha + w_t, t = 1, \ldots, T, \tag{8}$$

$$\hat{u}_t^2 - \hat{\sigma}^2 = z_t'\alpha + w_t, t = 1, \ldots, T, \tag{9}$$

$$|\hat{u}_t| = z_t'\alpha + w_t, t = 1, \ldots, T, \tag{10}$$

where $z_t = (1, z_{t2}, \ldots, z_{tm})'$ is a vector of m fixed regressors on which σ_t may depend, $\alpha = (\alpha_1, \ldots, \alpha_m)'$ and $w_t, t = 1, \ldots, T$, are treated as error terms. The Breusch–Pagan–Godfrey (BPG) LM criterion (Breusch and Pagan, 1979; Godfrey, 1978) may be obtained as the explained sum of squares (ESS) from the regression associated with (9) divided by $2\hat{\sigma}^4$. The Koenker (K) test statistic (Koenker, 1981) is T times the centered R^2 from regression (8). White's (W) test statistic is T times the centered R^2 from regression (8) using for z_t the $r \times 1$ observations on the non-redundant variables in the vector $x_t \otimes x_t$. These tests can be derived as LM-type tests against alternatives of the form H_A: $\sigma_t^2 = g(z_t'\alpha)$ where $g(\cdot)$ is a twice differentiable function. Under H_0 and standard asymptotic regularity conditions, $BPG \overset{asy}{\sim} \chi^2(m-1), K \overset{asy}{\sim} \chi^2(m-1), W \overset{asy}{\sim} \chi^2(r-1)$, where the symbol $\overset{asy}{\sim}$ indicates that the test statistic is asymptotically distributed as indicated (under H_0 as $T \to \infty$). The standard F statistic to test $\alpha_2 = \cdots = \alpha_m = 0$ in the context of (10) yields the Glejser (G) test (Glejser, 1969). Again, under H_0 and standard regularity conditions, $(T-k)G \overset{asy}{\sim} \chi^2(m-1)$. Below, we shall also consider $F(m-1, T-k)$ distribution as an approximation to the null distribution of this statistic. Honda (1988) has also suggested a size-correction formula [based on a general expansion given by Harris (1985)] for the BPG test. White's test was designed against the general alternative H_A. The above version of the Glejser test is valid for the special case where the variance is proportional to $z_t'\alpha$.

In Section 4, we describe a general solution to the problem of controlling the size of these tests. Our analysis further leads to two new results. First, in the context of the G test, Godfrey (1996) has recently shown that, unless the error distribution is symmetric, the test is deficient in the following sense. The residual-based test is not asymptotically equivalent to a conformable χ^2 test based on the true errors. Therefore, the G test may not achieve size control. We will show below that this difficulty is circumvented by our proposed MC version of the test. Second, we argue that from a MC test perspective, choosing the Koenker statistic rather than the BPG has no effect on size control.

3.1.2. Auxiliary regression tests against an unknown variance breakpoint

Tests against discrete breaks in variance at some specified date τ may be applied in the above framework by defining z_t as a dummy variable of the form $z_t = z_t(\tau)$, where

$$z_t(\tau) = \begin{cases} 0 & \text{if } t \leqslant \tau, \\ 1 & \text{if } t > \tau. \end{cases} \tag{11}$$

Typically, if the break-date τ is left unspecified and thus may take any one of the values $\tau = 1, \ldots, T-1$, a different test statistic may be computed for each one of these possible break-dates; see Pagan and Hall (1983). Here we provide a procedure to combine inference based on the resulting multiple tests, a problem not solved by Pagan and Hall (1983). Let BPG_τ be the BPG statistic obtained on using $z_t = z_t(\tau)$, where $\tau = 1, \ldots, T-1$. When used as a single test, the BPG_τ statistic is significant at level α when $BPG_\tau \geqslant \chi^2_\alpha(1)$, or equivalently when $\mathsf{G}_{\chi_1}(BPG_\tau) \leqslant \alpha$, where $\chi^2_\alpha(1)$ solves the equation $\mathsf{G}_{\chi_1}[\chi^2_\alpha(1)] = \alpha$ and $\mathsf{G}_{\chi_1}(x) = \mathsf{P}[\chi^2(1) \geqslant x]$ is the survival function of the $\chi^2(1)$ probability distribution. $\mathsf{G}_{\chi_1}(BPG_\tau)$ is the asymptotic p-value associated with BPG_τ. We propose here two methods for combining the BPG_τ tests.

The first one rejects H_0 when at least one of the p-values for $\tau \in J$ is sufficiently small, where J is some appropriate subset of the time interval $\{1, 2, \ldots, T-1\}$, such as $J = [\tau_1, \tau_2]$ where $1 \leqslant \tau_1 < \tau_2 \leqslant T-1$. In theory, J may be any non-empty subset of $\{1, 2, \ldots, T-1\}$. More precisely, we reject H_0 at level α when

$$pv_{\min}(BPG; J) \leqslant p_0(\alpha; J), \text{ where } pv_{\min}(BPG; J) \equiv \min\{\mathsf{G}_{\chi_1}(BPG_\tau) : \tau \in J\} \tag{12}$$

or, equivalently, when

$$F_{\min}(BPG; J) \geqslant F_{\min}(\alpha; J) \text{ where } F_{\min}(BPG; J) \equiv 1 - \min\{\mathsf{G}_{\chi_1}(BPG_\tau) : \tau \in J\}; \tag{13}$$

$p_0(\alpha; J)$ is the largest point such that $\mathsf{P}[pv_{\min}(BPG; J) \leqslant p_0(\alpha; J)] \leqslant \alpha$ under H_0, and $F_{\min}(\alpha; J) = 1 - p_0(\alpha; J)$. In general, to avoid over-rejecting, $p_0(\alpha; J)$ should be smaller than α. This method of combining tests was suggested by Tippett (1931) and Wilkinson (1951) in the case of independent test statistics. It is however clear that $BPG_\tau, \tau = 1, \ldots, T-1$, are not independent, with possibly a complex dependence structure.

The second method we consider consists in rejecting H_0 when the product (rather than the minimum) of the p-values $pv_\times(BPG; J) \equiv \prod_{\tau \in J} \mathsf{G}_\chi(BPG_\tau)$ is small, or equivalently when

$$F_\times(BPG; J) \geqslant \bar{F}_\times(J; \alpha), \text{ where } F_\times(BPG; J) \equiv 1 - \prod_{\tau \in J} \mathsf{G}_{\chi_1}(BPG_\tau); \tag{14}$$

$\bar{F}_\times(J; \alpha)$ is the largest point such that $\mathsf{P}[F_\times(BPG; J) \geqslant \bar{F}_\times(J; \alpha)] \leqslant \alpha$ under H_0. This general method of combining p-values was originally suggested by Fisher (1932) and Pearson (1933), again for independent test statistics.[8] We also propose here to consider a modified version of $F_\times(BPG; J)$ based on a subset of the p-values $\mathsf{G}_{\chi_1}(BPG_\tau)$.

[8] For further discussion of methods for combining tests, the reader may consult Folks (1984), Dufour (1989, 1990), Westfall and Young (1993), Dufour and Torrès (1998, 2000), and Dufour and Khalaf (2002a).

Specifically, we shall consider a variant of $F_{\times}(BPG; J)$ based on the m smallest p-values:

$$F_{\times}(BPG; \hat{J}_{(4)}) = 1 - \prod_{\tau \in \hat{J}_{(m)}} \mathsf{G}_{\chi_1}(BPG_\tau) \tag{15}$$

where $\hat{J}_{(m)}$ is the set of the m smallest p-values in the series $\{\mathsf{G}_{\chi_1}(BPG_\tau): \tau = 1, 2, \ldots, T-1\}$. The maximal number of p-values retained (m in this case) may be chosen to reflect (prior) knowledge on potential break dates; or as suggested by Christiano (1992), m may correspond to the number of local minima in the series $\mathsf{G}_{\chi_1}(BPG_\tau)$.

To derive exact tests based on $F_{\min}(BPG; J), F_{\times}(BPG; J)$ and $F_{\times}(BPG; \hat{J}_{(m)})$ is one of the main contributions of this paper. Indeed, their finite-sample and *even their asymptotic distributions* may be intractable. In Section 4, we show that the technique of MC tests provides a simple way of controlling their size. Our simulation experiments reported in Section 5 further illustrates their power properties.

3.2. Tests against ARCH-type heteroskedasticity

In the context of conditional heteroskedasticity, artificial regressions provide an easy way to compute tests for GARCH effects. Engle (1982) proposed an LM test based on (1) where

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \sigma_{t-i}^2 \varepsilon_{t-i}^2. \tag{16}$$

$\varepsilon_t|_{t-1} \sim \mathrm{N}(0,1)$ and $|_{t-1}$ denotes conditioning of information up to and including $t-1$. The hypothesis of homoskedasticity may then be formulated as H_0: $\alpha_1 = \cdots = \alpha_p = 0$. The Engle test statistic (which is denoted by E below) is given by TR^2, where T is the sample size, R^2 is the coefficient of determination in the regression of squared OLS residuals $\hat{u}_t^2$ on a constant and $\hat{u}_{t-i}^2$ $(i=1,\ldots,q)$. Under standard regularity conditions $E \overset{asy}{\sim} \chi^2(q)$. Lee (1991) has also shown that the same test is appropriate against GARCH(p,q) alternatives, i.e.

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \theta_i \sigma_{t-i}^2 + \sum_{i=1}^{q} \alpha_i \sigma_{t-i}^2 \varepsilon_{t-i}^2, \tag{17}$$

and the null hypothesis is H_0: $\alpha_1 = \cdots = \alpha_q = \theta_1 = \cdots = \theta_p = 0$. Lee and King (1993) proposed an alternative GARCH test which exploits the one-sided nature of H_A. The test statistic which is asymptotically standard normal under H_0 is

$$LK = \frac{\left\{(T-q)\sum_{t=q+1}^{T} [(\hat{u}_t/\hat{\sigma})^2 - 1)] \sum_{i=1}^{q} \hat{u}_{(t-i)}^2\right\} \Big/ \left\{\sum_{t=q+1}^{T} [(\hat{u}_t/\hat{\sigma})^2 - 1]^2\right\}^{1/2}}{\left\{(T-q)\sum_{t=q+1}^{T} \left(\sum_{i=1}^{q} \hat{u}_{t-i}^2\right)^2 - \left(\sum_{t=q+1}^{T}\left(\sum_{j=1}^{q} \hat{u}_{t-i}^2\right)\right)^2\right\}^{1/2}}. \tag{18}$$

Tests against ARCH-M heteroskedasticity (where the shocks affecting the conditional variance of y_t also have an effect on its conditional mean) can be applied in the context of

$$y_t = x_t'\beta + \sigma_t\phi + u_t, t = 1, \dots, T. \tag{19}$$

Bera and Ra (1995) show that the relevant LM statistic [against (16)] for given ϕ is:

$$LM(\phi) = \frac{1}{2+\phi^2}\hat{\gamma}'V\left[V'V - \frac{\phi^2}{2+\phi^2}V'X(X'X)^{-1}X'V\right]^{-1}V'\hat{\gamma}, \tag{20}$$

where $\hat{\gamma}$ is a $T\times 1$ vector with elements $\hat{\gamma}_t = [(\hat{u}_t/\hat{\sigma})^2 - 1] + \phi\hat{u}_t/\hat{\sigma}$ and V is a $T\times(q+1)$ matrix whose t-th row is $V_t = (1, \hat{u}_{t-1}^2, \dots, \hat{u}_{t-q}^2)$. In this case, under H_0, the parameter ϕ is unidentified. Bera and Ra (1995) also discuss the application of the Davies sup-LM test to this problem and show that this leads to more reliable inference. It is clear, however, that the asymptotic distribution required is quite complicated.

In Section 4, we describe a general solution to the problem of controlling the size of these tests. Our analysis further yields a new and notable result regarding the ARCH-M test: we show that the unidentified nuisance parameter is not a problem for implementing the MC version of the test. Indeed, it is easy to see that the statistic's finite sample null distribution is nuisance-parameter-free. The simulation experiment in Section 5.1 shows that this method works very well in terms of size and power. [9]

3.3. Tests based on grouping

An alternative class of tests assumes that observations can be ordered (e.g. according to time or some regressor) so that the variance is non-decreasing. Let $\hat{u}_{(t)}, t = 1, \dots, T$, denote the OLS residuals obtained after reordering the observations (if needed).

3.3.1. Goldfeld–Quandt tests against an unknown variance breakpoint

The most familiar test in this class is the Goldfeld and Quandt (1965, GQ) test which involves separating the ordered sample into three subsets and computing separate OLS regressions on the first and last data subsets. Let $T_i, i = 1, 2, 3$, denote the number of observations in each of these subsets ($T = T_1 + T_2 + T_3$). The test statistic, which is $F(T_3 - k, T_1 - k)$ distributed under (1)–(6) and H_0, is

$$GQ(T_1, T_3, k) = \frac{S_3/(T_3 - k)}{S_1/(T_1 - k)} \tag{21}$$

where S_1 and S_3 are the sum of squared residuals from the first T_1 and the last T_3 observations ($k < T_1$ and $k < T_3$). The latter distributional result is exact provided the ranking index does not depend on the parameters of the constrained model. Setting $\mathsf{G}_{F(T_3-k,T_1-k)}(x) = \mathsf{P}[F(T_3 - k, T_1 - k) \geqslant x]$, we denote $pv[GQ; T_1, T_3, k] = \mathsf{G}_{F(T_3-k,T_1-k)}[GQ(T_1, T_3, k)]$ the p-value associated with $GQ(T_1, T_3, k)$.

[9] Demos and Sentana (1998) proposed one-sided tests for ARCH. Similarly, Beg et al. (2001) introduced a one-sided sup-type generalization of the Bera–Ra test, together with simulation-based cut-off points, because of the intractable asymptotic null distributions involved. For further discussion of these difficulties, see also Andrews (2001). The MC test method should also be useful with these procedures.

The GQ test is especially relevant in testing for breaks in variance.[10] Here we extend this test to account for an unknown (or unspecified) break-date. We propose (as for the BPG test) statistics of the form:

$$F_{\min}(GQ; K) \equiv 1 - \min\{pv[GQ; T_1, T_3, k]: (T_1, T_3) \in K\}, \tag{22}$$

$$F_{\times}(GQ; K) \equiv 1 - \prod_{(T_1,T_3)\in K} pv(GQ; T_1, T_3, k], \tag{23}$$

where K is any appropriate non-empty subset of

$$K(k, T) = \{(T_1, T_3) \in \mathbb{Z}^2: k + 1 \leqslant T_1 \leqslant T - k - 1 \text{ and } k + 1 \leqslant T_3 \leqslant T - T_1\}, \tag{24}$$

the set of the possible subsample sizes compatible with the definition of the GQ statistic. Reasonable choices for K could be $K = S_1(T, T_2, L_0, U_0)$ with

$$S_1(T, T_2, L_0, U_0) \equiv \{(T_1, T_3): L_0 \leqslant T_1 \leqslant U_0 \text{ and } T_3 = T - T_1 - T_2 \geqslant 0\}, \tag{25}$$

where T_2 represents the number of central observations while L_0 and U_0 are minimal and maximal sizes for the subsamples ($0 \leqslant T_2 \leqslant T - 2k - 2, L_0 \geqslant k + 1, U_0 \leqslant T - T_2 - k - 1$), or

$$K = S_2(T, L_0, U_0) = \{(T_1, T_3): L_0 \leqslant T_1 = T_3 \leqslant U_0\} \tag{26}$$

where $L_0 \geqslant k + 1$ and $U_0 \leqslant I[T/2]; I[x]$ is the largest integer less than or equal to x. According to definition (25), $\{GQ(T_1, T_3, k): (T_1, T_3) \in K\}$ defines a set of GQ statistics, such that the number T_2 of central observations is kept constant (although the sets of the central observations differ across the GQ statistics considered); with (26), $\{GQ(T_1, T_3, k): (T_1, T_3) \in K\}$ leads to GQ statistics such that $T_1 = T_3$ (hence with different numbers of central observations). As with the BPG statistics, we also consider

$$F_{\times}(GQ; \hat{K}_{(m)}) \equiv 1 - \prod_{(T_1,T_3)\in \hat{K}_{(m)}} pv[GQ; T_1, T_3, k],$$

where $\hat{K}_{(m)}$ selects the m smallest p-values from the set $\{pv[GQ; T_1, T_3, k]: (T_1, T_3) \in K\}$.

It is clear the null distribution of these statistics may be quite difficult to obtain, *even asymptotically*. In this regard, this paper makes the following contribution: we show in Section 4 that the level of a test procedure based on any one of these statistics can be controlled quite easily by using the MC version of these tests. Our simulations, reported in Section 5, further show that our proposed tests perform quite well in terms of power.

3.3.2. Generalized Bartlett tests

Under the Gaussian assumption (6), the likelihood ratio criterion for testing H_0 against H_4 is a (strictly) monotone increasing transformation of the statistic:

$$LR_{(H_4)} = T \ln(\hat{\sigma}^2) - \sum_{i=1}^{p} T_i \ln(\hat{\sigma}_i^2), \tag{27}$$

[10] Pagan and Hall (1983, p. 177) show that the GQ test for a break in variance and the relevant dummy-variable based BPG test are highly related.

where $\hat{\sigma}^2$ is the ML estimator (assuming i.i.d. Gaussian errors) from the pooled regression (1) while $\hat{\sigma}_i^2, i = 1, \ldots p$, are the ML estimators of the error variances for the p subgroups (which, due to the common regression coefficients require an iterative estimation procedure). If one further allows the regression coefficient vectors to differ between groups (under both the null and the alternative hypothesis), one gets the extension to the linear regression setup of the well-known Bartlett (1937) test for variance homogeneity.[11] Note Bartlett (1937) studied the special case where the only regressor is a constant, which is allowed to differ across groups. Other (quasi-LR) variants of the Bartlett test, involving degrees of freedom corrections or different ways of estimating the group variances, have also been suggested; see, for example, Binkley (1992).

In the context of H_2 Ramsey (1969) suggested a modification to Bartlett's test that can be run on BLUS residuals from ordered observations. Following Griffiths and Surekha (1986), we consider an OLS-based version of Ramsey's test which involves separating the residuals $\hat{u}_{(t)}, t = 1, \ldots, T$, into three disjoint subsets G_i with $T_i, i = 1, 2, 3$, observations, respectively. The test statistic which is asymptotically $\chi^2(2)$ under H_0 is:

$$RB = T\ln(\hat{\sigma}^2) - \sum_{i=1}^{3} T_i \ln(\hat{\sigma}_i^2), \quad \hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_{(t)}^2, \quad \hat{\sigma}_i^2 = \frac{1}{T_i}\sum_{t\in G_i} \hat{u}_{(t)}^2. \tag{28}$$

3.3.3. Szroeter-type tests

Szroeter (1978) introduced a wide class of tests based on statistics of the form

$$\tilde{h} = \left(\sum_{t\in A} h_t \tilde{u}_t^2\right) \Big/ \left(\sum_{t\in A} \tilde{u}_t^2\right), \tag{29}$$

where A is some non-empty subset of $\{1, 2, \ldots, T\}$, the $\tilde{u}_t$'s are a set of residuals, and the h_t's are a set of non-stochastic scalars such that $h_s \leqslant h_t$ if $s < t$. Szroeter suggested several special cases (obtained by selecting different weights h_t), among which we consider the following [based on the OLS residuals from a single regression, i.e. $\tilde{u}_t = \hat{u}_{(t)}$]:

$$SKH = \left[\sum_{t=1}^{T} 2\left[1 - \cos\left(\frac{\pi t}{t+1}\right)\right] \hat{u}_{(t)}^2\right] \Big/ \left(\sum_{t=1}^{T} \hat{u}_{(t)}^2\right), \tag{30}$$

$$S_N = \left(\frac{6T}{T^2 - 1}\right)^{1/2} \left(\frac{\sum_{t=1}^{T} t\hat{u}_{(t)}^2}{\sum_{t=1}^{T} \hat{u}_{(t)}^2} - \frac{T+1}{2}\right), \tag{31}$$

$$S_F = \left(\sum_{t=T_1+T_2+1}^{T} \hat{u}_{(t)}^2\right) \Big/ \left(\sum_{t=1}^{T_1} \hat{u}_{(t)}^2\right) \equiv S_F(T_1, T - T_1 - T_2). \tag{32}$$

[11] In this case, the estimated variances are $\hat{\sigma}_i^2 = S_i/T_i, i = 1, \ldots, p$, and $\hat{\sigma}^2 = \sum_{i=1}^{p} T_i S_i/T$, where S_i is the sum of squared errors from a regression which only involves the observations in the ith group. This, of course, requires one to use groups with sufficient numbers of observations.

Under the null hypothesis, S_N follows a $\mathrm{N}(0,1)$ distribution asymptotically. Exact critical points for SKH [under (6)] may be obtained using the Imhof method. Szroeter recommends the following bounds tests. Let h_L^* and h_U^* denote the bounds for the Durbin and Watson (1950) test corresponding to $T+1$ observations and k regressors. Reject the homoskedasticity hypothesis if $SKH > 4 - h_L^*$, accept if $SKH < 4 - h_u^*$, and otherwise treat the model as inconclusive. King (1981) provided revised bounds for use with SKH calculated from data sorted such that, under the alternative, the variances are non-increasing. Harrison (1980, 1981, 1982), however, showed there is a high probability that the Szroeter and King bounds tests will be inconclusive; in view of this, he derived and tabulated beta-approximate critical values based on the Fisher distribution.

As with the GQ test, the Szroeter's S_F statistic may be interpreted as a variant of the GQ statistic where the residuals from separate regressions have been replaced by those from the regression based on all the observations, so that S_3 is replaced by $\tilde{S}_3 = \sum_{t=T_1+T_2+1}^{T} \hat{u}_{(t)}^2$ and S_1 by $\tilde{S}_1 = \sum_{t=1}^{T_1} \hat{u}_{(t)}^2$. Harrison and McCabe (1979) suggested a related test statistic based on the ratio of the sum of squares of a subset of $\{\hat{u}_{(t)}, t = 1, \ldots, T\}$, to the total sum of squares:

$$HM = \left(\sum_{t=1}^{T_1} \hat{u}_{(t)}^2 \right) \Big/ \left(\sum_{t=1}^{T} \hat{u}_{(t)}^2 \right), \tag{33}$$

where $T_1 = I[T/2]$. Although the test critical points may also be derived using the Imhof method, Harrison and McCabe proposed the following bounds test. Let $b_L^* = [1 + b(T - T_1, T - k, T - k)]^{-1}$ and $b_U^* = [1 + b(T - T_1 - k, T_1, T - k)]^{-1}$, where $b(v_1, v_2, v_3) = (v_1/v_3)F_\alpha(v_1, v_2)$ and $F_\alpha(v_1, v_2)$ refers to the level α critical value from the $F(v_1, v_2)$ distribution. H_0 is rejected if $HM < b_L^*$, it is accepted if $HM > b_U^*$, and otherwise the test is inconclusive. Beta approximations to the null distribution of the HM statistic have also been suggested, but they appear to offer little savings in computational cost over the exact tests (see Harrison, 1981).

McCabe (1986) proposed a generalization of the HM test to the case of heteroskedasticity occurring at unknown points. The test involves computing the maximum HM criterion over several sample subgroups (of size m). The author suggests Bonferroni-based significance points using the quantiles of the Beta distribution with parameters $[m/2, (t - m - k)/2]$. McCabe discusses an extension to the case where m is unknown. The proposed test is based on the maximum of the successive differences of the order statistics and also uses approximate beta critical points.

In this context, our contribution is twofold. First, we show in Section 4 that exact MC versions of these tests can be easily obtained even in non-Gaussian context. Secondly, our simulation results reveal that MC Szroeter-type tests have *definite* power advantages over the MC versions of commonly used tests such as the BPG test. This observation has noteworthy implications for empirical practice, since it seems that in spite of the many available homoskedasticity tests, practitioners (see Table 1) seem to favor BPG-type tests.

3.3.4. Generalized Cochran–Hartley tests

Cochran (1941) and Hartley (1950) proposed two classic tests against grouped heteroskedasticity (henceforth denoted C and H, respectively) in the context of simple Gaussian location-scale models (i.e., regressions that include only a constant). These are based on maximum and minimum subgroup error variances. Extensions of these tests to the more general framework of linear regressions have been considered by Rivest (1986). The relevant statistics then take the form:

$$C = \max_{1 \leqslant i \leqslant p} (s_i^2) \Big/ \sum_{i=1}^{p} s_i^2, \quad H = \max_{1 \leqslant i \leqslant p} (s_i^2) / \min_{1 \leqslant i \leqslant p} (s_i^2) \tag{34}$$

where s_i^2 is the unbiased error variance estimator from the ith separate regression ($1 \leqslant i \leqslant p$). Although critical values have been tabulated for the simple location-scale model [see Pearson and Hartley (1976, pp. 202–203)], these are not valid for more general regression models, and Rivest (1986) only offers an asymptotic justification.

In this regard, this paper makes two contributions. We first show that the classic Cochran and Hartley tests can easily be implemented as finite-sample MC tests in the context of the regression model (1)–(4). Secondly, we introduce variants of the Cochran and Hartley tests that may be easier to implement or more powerful than the original procedures. Specifically, we consider replacing, in the formula of these statistics, the residuals from separate regressions by the OLS residuals from the pooled regression (1), possibly after the data have been resorted according to some exogenous variable. This will reduce the loss in degrees of freedom due to the separate regressions. The resulting test statistics will be denoted C_r and HR_r respectively. Clearly, standard distributional theory does not apply to these modified test criteria, but they satisfy the conditions required to implement them as MC tests. These results are new and have constructive implications for empirical practice; indeed, our simulations (reported in Section 5) show that such tests tend to perform well relative to the LR-type test presented above.

3.3.5. Grouping tests against a mean-dependent variance

Most of the tests based on grouping, as originally suggested, are valid for alternatives of the form H_2. A natural extension to alternatives such as H_3 involves sorting the data conformably with $\hat{y}_t$. However this complicates the finite-sample distributional theory; see Pagan and Hall (1983) or Ali and Giaccotto (1984). Here we propose the following solution to this problem. Whenever the alternative tested requires ordering the sample following the fitted values of a preliminary regression, rather than sorting the data, sort the residuals $\hat{u}_t, t = 1, \ldots, T$, following $\hat{y}_t$ and proceed. Provided the fitted values $(\hat{y}_1, \ldots, \hat{y}_T)'$, are independent of the least-squares residuals $(\hat{u}_1, \ldots, \hat{u}_T)'$ under the null hypothesis, as occurs for example under the Gaussian assumption (6), this will preserve the pivotal property of the tests and allow the use of MC tests. Note finally that this simple modification [not considered by Pagan and Hall (1983) or Ali and Giaccotto (1984)] solves a complicated distributional problem, underscoring again the usefulness of the MC test method in this context.

4. Finite-sample distributional theory

We will now show that all the statistics described in Section 3 have null distributions which are free of nuisance parameters and show how this fact can be used to perform a finite-sample MC test of homoskedasticity using any one of these statistics. For that purpose, we shall exploit the following general proposition.

Proposition 1 (Characterization of pivotal statistics). *Under the assumptions and notations* (1)–(2), *let* $S(y,X) = (S_1(y,X), S_2(y,X), \ldots, S_m(y,X))'$ *be any vector of real-valued statistics* $S_i(y,X), i = 1,\ldots,m$, *such that*

$$S(cy + Xd, X) = S(y,X), \text{ for all } c > 0 \text{ and } d \in \mathbb{R}^k. \tag{35}$$

Then, for any positive constant $\sigma_0 > 0$, *we can write*

$$S(y,X) = S(u/\sigma_0, X), \tag{36}$$

and the conditional distribution of $S(y,X)$, *given* X, *is completely determined by the matrix* X *and the conditional distribution of* $u/\sigma_0 = \Delta\varepsilon/\sigma_0$ *given* X, *where* $\Delta =$ diag $(\sigma_t : t = 1,\ldots,T)$. *In particular, under* H_0 *in* (4), *we have*

$$S(y,X) = S(\varepsilon, X) \tag{37}$$

where $\varepsilon = u/\sigma$, *and the conditional distribution of* $S(y,X)$, *given* X, *is completely determined by the matrix* X *and the conditional distribution of* ε *given* X.

Proof. The result follows on taking $c = 1/\sigma_0$ and $d = -\beta/\sigma_0$, which entails, by (1),

$$cy + Xd = (X\beta + u)/\sigma_0 - X\beta/\sigma_0 = u/\sigma_0.$$

Then, using (35), we get (36), so the conditional distribution of $S(y,X)$ only depends on X and the conditional distribution of u/σ_0 given X. The identity $u_0 = \Delta\varepsilon$ follows from (2). Finally, under H_0 in (4), we have $u = \Delta\varepsilon = \sigma\varepsilon$, hence, on taking $\sigma_0 = \sigma$, we get $u/\sigma_0 = \varepsilon$ and $S(y,X) = S(\varepsilon,X)$. □

It is of interest to note that (36) holds under both the general heteroskedastic model (1)–(2) and the homoskedastic model obtained by imposing (4), without any parametric distributional assumption on the disturbance vector u [such as (3)]. Then, assuming (3), we see that the conditional distribution of $S(y,X)$, given X, is free of nuisance parameters and thus may be simulated. Of course, the same will hold for any transformation of the components of $S(y,X)$, such as statistics defined as the supremum or the product of several statistics (or p-values).

It is relatively easy to check that all the statistics described in Section 3 satisfy the invariance condition (35). In particular, on observing that model (1)–(2) and the hypothesis (4) are invariant to general transformations of y to $y_* = cy + Xd$, where $c > 0$ and $d \in \mathbb{R}^k$, on y, it follows that LR test statistics against heteroskedasticity, such the Bartlett test based on $LR_{(\mathrm{H}_4)}$ in (27), satisfy (35) (see Dagenais and Dufour, 1991; and Dufour and Dagenais, 1992), and so have null distributions which are free of nuisance parameters. For the other statistics, the required results follow on observing that they are scale-invariant functions of OLS residuals. For that purpose, it will be useful to state the following corollary of Proposition 1.

Corollary 2 (Pivotal property of residual-based statistics). *Under the assumptions and notations (1)–(2), let $S(y,X) = (S_1(y,X), S_2(y,X), \ldots, S_m(y,X))'$ be any vector of real-valued statistics $S_i(y,X), i = 1,\ldots,m$, such that $S(y,X)$ can be written in the form*

$$S(y,X) = \bar{S}(A(X)y,X), \tag{38}$$

where $A(X)$ is any $n \times k$ matrix ($n \geqslant 1$) such that

$$A(X)X = 0 \tag{39}$$

and $\bar{S}(A(X)y,X)$ satisfies the scale-invariance condition

$$\bar{S}(cA(X)y,X) = \bar{S}(A(X)y,X), \quad \text{for all } c > 0. \tag{40}$$

Then, for any positive constant $\sigma_0 > 0$, we can write

$$S(y,X) = \bar{S}(A(X)u/\sigma_0,X) \tag{41}$$

and the conditional distribution of $S(y,X)$, given X, is completely determined by the matrix X jointly with the conditional distribution of $A(X)u/\sigma_0$ given X. In particular, under H_0 in (4), we have $S(y,X) = \bar{S}(A(X)y,X)$, where $\varepsilon = u/\sigma$, and the conditional distribution of $S(y,X)$, given X, is completely determined by the matrix X and the conditional distribution of $A(X)\varepsilon$ given X.

It is easy to see that the invariance conditions (38)–(40) are satisfied by any scale-invariant function of the OLS residuals from the regression (1), i.e. any statistic of the form $S(y,X) = \bar{S}(\hat{u},X)$ such that $\bar{S}(c\hat{u},X) = \bar{S}(\hat{u},X)$ for all $c > 0$ [in this case, we have $A(X) = I_T - X(X'X)^{-1}X'$]. This applies to all the tests based on auxiliary regressions described in Section 3.1 as well as the tests against ARCH-type heteroskedasticity (Section 3.2). On the other hand, the tests designed against grouped heteroskedasticity (Section 3.3) involve residuals from subsets of observations. These also satisfy the sufficient conditions of Corollary 2 although the $A(X)$ matrix involved is different. For example, for the GQ statistic, we have:

$$A(X) = \begin{bmatrix} M(X_1) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & M(X_3) \end{bmatrix}, \tag{42}$$

where $M(X_i) = I_{T_i} - X_i(X_i'X_i)^{-1}X_i', X = [X_1', X_2', X_3']'$ and X_i is a $T_i \times k$ matrix. The number n of rows in $A(X)$ can be as large as one wishes so several regressions of this type can be used to compute the test statistic, as done for the combined GQ statistic [see (22)]. Finally, the required invariance conditions are also satisfied by statistics built on other types of residuals, such as residuals based on least absolute deviation (instead of least squares) and various M-estimators. Studying such statistics would undoubtedly be of interest especially for dealing with nonnormal (possibly heavy-tailed) distributions. However, in view of the statistics considered in Section 3, we shall restrict ourselves in the sequel to statistics based on least squares methods.

Let us now make the parametric distributional assumption (3). Then we can proceed as follows to perform a finite-sample test based on any statistic, say $S_0 = S(y, X)$, whose null distribution (given X) is free of nuisance parameters. Let $G(x)$ be the survival function associated with S_0 under H_0, i.e. we assume $G : \mathbb{R} \to [0, 1]$ is the function such that $G(x) = \mathsf{P}_{\mathrm{H}_0}[S_0 \geqslant x]$ for all x, where $\mathsf{P}_{\mathrm{H}_0}$ refers to the relevant probability measure (under H_0). When the distribution of S_0 is *continuous*, we have $G(x) = 1 - F(x)$ where $F(x) = \mathsf{P}_{\mathrm{H}_0}[S_0 \leqslant x]$ is the distribution function of S_0 under H_0. Without loss of generality, we consider a right-tailed procedure: H_0 rejected at level α when $S_0 > c(\alpha)$, where $c(\alpha)$ is the appropriate critical value such that $G[c(\alpha)] = \alpha$, or equivalently (with probability 1) when $G(S_0) \leqslant \alpha$ [i.e. when the p-value associated with the observed value of the statistic is less than or equal to α].

Now suppose we can generate N i.i.d. replications of the error vector ε according to (3). This leads to N simulated samples and N independent realizations of the test statistic $S_1, \ldots, S_N$. The associated MC critical region is

$$\hat{p}_N(S_0) \leqslant \alpha, \tag{43}$$

$$\hat{p}_N(x) = \frac{N\hat{G}_N(x) + 1}{N + 1}, \quad \hat{G}_N(x) = \frac{\sum_{i=1}^{N} \mathbf{1}_{[0,\infty)}(S_i - x)}{N},$$

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

Then, provided the distribution function of S_0 induced by $\mathsf{P}_{\mathrm{H}_0}$ under H_0 is continuous,

$$\mathsf{P}_{\mathrm{H}_0}[\hat{p}_N(S_0) \leqslant \alpha] = \frac{I[\alpha(N+1)]}{N+1}, \text{for } 0 \leqslant \alpha \leqslant 1. \tag{44}$$

Note that the addition of 1 in the numerator and denominator of $\hat{p}_N(x)$ is important for (44) to hold. In particular, if N is chosen so that $\alpha(N + 1)$ is an integer, we have $\mathsf{P}_{\mathrm{H}_0}[\hat{p}_N(S_0) \leqslant \alpha] = \alpha$; see Dufour and Kiviet (1998). Thus the critical region (43) has the same size as the critical region $G(S_0) \leqslant \alpha$. The MC test so obtained is theoretically exact, irrespective of the number N of replications used.[12] Note that the procedure is closely related to the parametric bootstrap, with however a fundamental difference: bootstrap tests are, in general, provably valid for $N \to \infty$. See Dufour and Kiviet (1996, 1998), Kiviet and Dufour (1997), Dufour and Khalaf (2001, 2002a, b, 2003) for some econometric applications of MC tests. Finally, it is clear from the statement of Assumption (3) that normality does not play any role for the validity of the MC

[12] For example, for $\alpha = 0.05$, N can be as low as $N = 19$. Of course, this does not mean that a larger number of replications is not preferable, for raising N will typically increase the test power and decrease its sensitivity to the underlying randomization. However the simulation results reported below suggest that increasing N beyond 99 only has a small effect on power.

procedure just described. So we can consider in this way alternative error distributions such as heavy-tailed distributions like the Cauchy distribution.[13]

5. Simulation experiments

In this section, we present simulation results illustrating the performance of the procedures described in the preceding sections. We consider the most popular heteroskedastic alternatives studied in this literature [for references, refer to Section 1]: (1) GARCH-type heteroskedasticity; (2) variance as a linear function of exogenous variables; (3) grouped heteroskedasticity; (4) variance break at a (possibly unspecified) point. Our designs are carefully modelled after several well known power studies [including: Honda (1988), Binkley (1992), Godfrey (1996), Bera and Ra (1995) and Lumsdaine (1995)] to make our results comparable with published work.[14]

5.1. Tests for ARCH and GARCH effects

For ARCH and GARCH alternatives, our simulation experiment was based on (17), (19), with $q = 1, T = 50, 100, k = I[T^{1/2}] + 1, \beta = (1, 1, \ldots, 1)'$ and $\alpha_0 = 1$. Four experiments were conducted with parameters set as follows, to make the results of our study comparable with those obtained by Lee and King (1993), Lumsdaine (1995) and Bera and Ra (1995): (i) $\phi = \alpha_1 = \theta_1 = 0$; (ii) $\phi = \theta_1 = 0, \alpha_1 = 0.1, 0.5, 0.9, 1$; (iii) $\phi = 0, (\alpha_1, \theta_1) = (0.1, 0.5), (0.25, 0.65), (0.4, 0.5), (0.15, 0.85), (0.05, 0.95)$; (iv) $\theta_1 = 0, \phi = -2, \alpha_1 = 0, 0.1, 0.9$. Note that some combinations fall on the boundary of the region $\alpha_1 + \theta_1 \leqslant 1$. In experiments (i)–(iii), $\varepsilon_t \sim \mathrm{N}(0, 1)$. In experiment (iv), we considered alternative error distributions, according to the examples studied by Godfrey (1996): $\mathrm{N}(0, 1), \chi^2(2), t(5)$ and *Cauchy*. The regressors were generated as i.i.d. according to a $\mathrm{U}(0, 10)$ distribution and kept constant over each individual experiment. In the case of experiment (iv), we also considered an alternative regressor set, obtained by drawing (independently) from a Cauchy distribution (centered and re-scaled conformably with the previous design). For further reference, we shall denote by D1 the *uniform*-based design D1 and by D2 the *Cauchy*-based design. Both D1 and D2 include a constant regressor. The Engle and Lee-King tests were applied in all cases. In experiment (iv), we also applied the Bera-Ra sup-LM test (see Section 3.2 for formulae and related references), in which case we have only computed MC versions of the tests. The MC

[13] As already pointed out, the only tests for which normality may play a central role in order to control size are those designed against a variance which is a function of the mean and where the least squares (LS) residuals are sorted according to the LS fitted values $\hat{y}_t, t = 1, \ldots, T$. Since the distribution of the latter depends on nuisance parameters (for example, the regression coefficients β), it is not clear that a test statistic which depends on both $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_T)'$ and $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T)'$ will have a null distribution free of nuisance parameters under the general distributional assumption (3).

[14] The MC studies we consider assess the various tests, assuming a correctly specified model. We do not address the effects on the tests which result from misspecifying the model and/or the testing problem. For space considerations, we report in the following a number of representative findings. More detailed results are presented in Dufour et al. (2001).

Table 2
Tests against ARCH and GARCH heteroskedasticity: normal error distribution and D1 design

99 MC reps.		$T=50$				$T=100$			
		E		LK		E		LK	
	$(\phi, \alpha_1, \theta_1)$	ASY	MC	ASY	MC	ASY	MC	ASY	MC
H_0	(0,0,0)	3.06	4.94	4.04	5.01	3.63	5.18	4.72	5.22
ARCH	(0,0.1,0)	6.42	8.60	9.67	**11.31**	11.83	13.61	17.01	**17.22**
	(0, 0.5, 0)	31.56	35.68	39.96	**42.28**	64.18	66.43	71.93	**71.54**
	(0, 0.9, 0)	50.57	54.76	58.89	**60.71**	84.38	85.82	88.99	**88.97**
	(0, 1, 0)	53.43	57.77	61.82	**63.61**	86.40	87.50	90.60	**90.24**
GARCH	(0, 0.1, 0.5)	6.89	9.16	10.45	**12.05**	12.54	14.39	17.89	**18.35**
	(0, 0.25, 0.65)	16.26	19.92	23.43	**25.20**	38.36	40.74	46.93	**47.29**
	(0, 0.40, 0.50)	26.12	30.25	34.48	**36.51**	57.44	59.65	65.94	**65.69**
	(0, 0.15, 0.85)	13.45	16.79	19.96	**21.70**	28.97	31.04	37.11	**37.93**
	(0, 0.05, 0.95)	10.02	12.77	15.37	**17.05**	18.17	20.15	25.92	**26.28**

Table 3
Size and power of MC ARCH-M tests: various error distributions, D2 design

99 MC reps.		$T=50$					$T=100$				
$\phi=-2$		E		LK		BR	E		LK		BR
α_1	Error	ASY	MC	ASY	MC	MC	ASY	MC	ASY	MC	MC
0	N(0,1)	3.28	4.82	5.19	4.83	5.16	3.55	5.15	5.60	5.31	5.36
	$\chi^2(2)$	2.86	5.12	4.32	5.13	5.29	3.16	4.79	4.80	4.82	5.04
	$t(5)$	2.20	5.10	2.88	5.26	4.96	2.35	4.82	3.01	4.87	4.90
	Cauchy	1.73	4.97	2.30	5.01	5.17	1.62	5.27	2.12	5.23	5.23
0.1	N(0,1)	8.79	11.55	14.10	13.18	**14.52**	16.95	18.80	24.02	22.24	**25.01**
	$\chi^2(2)$	31.38	40.97	40.29	42.74	**48.42**	54.44	60.07	61.48	60.88	**66.50**
	$t(5)$	21.11	44.90	32.08	48.63	**58.36**	48.88	66.65	57.40	68.28	**74.01**
	Cauchy	45.23	65.63	53.15	69.88	**73.76**	65.98	82.32	71.01	84.77	**87.42**
0.9	N(0,1)	39.79	43.16	47.77	45.96	**62.65**	63.00	64.64	68.47	66.75	**83.26**
	$\chi^2(2)$	27.99	34.37	32.89	35.71	**55.27**	37.62	42.78	43.22	43.62	**68.35**
	$t(5)$	26.88	47.77	36.51	50.70	**57.09**	42.96	60.47	51.05	62.15	**70.09**
	Cauchy	53.12	73.68	61.13	77.29	**80.86**	70.13	84.72	74.70	86.79	**89.41**

tests were implemented with $N = 99$ replications. In the case of experiment (iv), we also considered $N = 199$, 499 and 999.

Tables 2 and 3 report rejection percentages for a nominal level of 5%; 10000 replications were considered; in these tables (as well as later ones), the figures associated with best performing exact procedures in terms of power (under the alternatives) are set in bold face characters. In general, the most notable observation is that the Engle test is undersized, even with $T=100$, which can lead to substantial power losses. This is in accordance with the results of Lee and King (1993) and several references cited there. Although undersize problems are evident under D1 and normal errors, more serious size distortions are observed with $\chi^2(2)$, $t(5)$ and Cauchy errors. The size of the Lee–King test is better than that of the Engle test but is still below the nominal level particularly with non-normal errors.

These results show that MC tests yield noticeable effective power gains, even with uniform designs and normal errors. In the case of $\chi^2(2)$, $t(5)$ and Cauchy errors, improvements in power are quite substantial (such as a 50% increase with $T=50$). As emphasized in Bera and Ra (1995), power improvements are especially important for ARCH and GARCH tests since failing to detect conditional variance misspecifications leads to inconsistencies in conditional moment estimates. The Lee–King MC test is always more powerful than the Engle test. For ARCH-M alternatives, there is a substantial power gain from using the sup-LM MC test. It is also worth noting that possible problems at boundary parameter values were not observed. The power advantage of the MC sup-LM along with the documented difficulties regarding the Davies sup-LM test, makes the MC Bera–Ra test quite attractive. [15] Further, these results show clearly that the MC test provides a straightforward finite-sample solution to the problem of unidentified nuisance parameters.

Experiment (iv) allows to assess the effect of increasing MC replications. Our results (reported in Table 4) show that effective power improvements are not noticeable beyond $N=99$; indeed, power increases with N albeit quite modestly. This result, which agrees with available evidence from the bootstrap literature (see for example Davison and Hinkley, 1997, p. 143), confirms that our simulation studies with 99 replications are indeed representative.

5.2. Tests of variance as a linear function of exogenous variables

The model used is (1) with $T=50,100$ and $k=6$. [16] The regression coefficients were set to one. The following specification for the error variance were considered: (i) $\sigma_t^2=x_t'\alpha$, $t=1,\ldots,T$, where $\alpha=(1,0,\ldots,0)'$ under H_0, and $\alpha=(1,1,\ldots,1)'$ under H_A, and (ii) $\sigma_t^2=\alpha_0+\alpha_1 x_{2t}$, $t=1,\ldots,T$, where $\alpha_0=1$, $\alpha_1=0$, under H_0 and $\alpha_0=0$, $\alpha_1=1$, under H_A. The former specification implies that the variance is a linear function of $E(Y_t)$ and the latter is the case where the variance is proportional to one regressor. The regressors are generated as $U(0,10)$. The tests examined are: the Goldfeld–Quandt (GQ) test [see (21)], with $T_2=T/5$, $T_1=T_3=(T-T_2)/2$; the Breusch–Pagan–Godfrey (BPG) test [see (8)], based on the asymptotic distribution (ASY) or using the size correction formula (BRT) proposed by Honda (1988, Section 2); Koenker's (K) test [see (9)]; White's (W) test [see (8)]; Glejser's (G) test based on (10); Ramsey's version of Bartlett's test (RB) [see (28)], with $T_1=T_3=I[T/3]$ and $T_2=T-(T_1+T_3)$;

[15] This also suggests that an MC version of the Beg et al. (2001) one-sided test for ARCH-M may also result in power improvements.

[16] Tables of critical points required for the Szroeter's tests are available for $n\leqslant 100$ and $k\leqslant 6$; see King (1981) and Harrison (1982).

Table 4
Power of MC ARCH-M tests: increasing MC replications

$\phi=-2$		$T=50$						$T=100$					
		$\alpha_1=0.1$			$\alpha_1=0.9$			$\alpha_1=0.1$			$\alpha_1=0.9$		
	Reps	*E*	*LK*	*BR*	*E*	*LK*	*BR*	*E*	*LK*	*BR*	*E*	*LK*	*BR*
N	99	11.55	13.18	14.52	43.16	45.96	62.52	18.80	22.24	25.01	64.64	66.75	83.26
	199	11.61	13.45	14.76	43.74	46.63	63.12	19.19	22.56	25.33	64.75	67.01	83.5
	499	11.92	13.62	15.03	44.17	46.77	63.33	19.40	22.75	25.64	64.93	67.27	83.78
	999	11.88	13.56	14.89	44.22	47.16	63.44	19.39	22.63	25.63	64.89	67.38	83.88
χ^2	99	40.97	42.74	48.42	34.37	35.71	55.27	60.07	60.88	66.50	42.78	43.62	68.35
	199	41.49	43.20	49.64	34.75	35.92	55.68	61.14	62.00	67.74	43.42	44.01	68.86
	499	42.53	44.30	50.35	34.85	35.98	56.36	61.82	62.40	68.48	43.49	43.81	69.31
	999	42.59	44.38	50.71	35.06	35.92	56.39	61.68	62.31	68.57	43.73	44.02	69.49
C	99	65.63	69.88	73.76	73.68	77.29	80.86	82.32	84.77	87.42	84.72	86.79	89.41
	199	66.89	71.10	75.10	74.95	78.83	82.17	83.42	85.97	88.74	85.63	87.99	90.41
	499	67.49	72.14	76.10	75.66	79.74	82.66	83.92	86.69	89.62	86.25	88.43	90.78
	999	67.83	72.43	76.62	75.75	79.71	82.87	84.02	86.85	89.68	86.28	88.51	91.23
t	99	44.90	48.63	58.36	47.74	50.70	57.09	66.65	68.28	74.01	60.47	62.15	70.09
	199	46.57	49.93	60.17	48.94	51.76	58.44	68.93	70.31	76.71	62.54	63.78	71.68
	499	47.99	51.00	61.68	50.25	52.54	59.39	70.14	71.34	77.86	63.96	64.88	72.64
	999	48.22	51.15	62.12	50.78	53.14	59.77	70.99	71.93	78.12	64.18	65.21	72.90

Szroeter's S_F test [see (32)], where for convenience, T_1 and T_2 are set as in the GQ test; Szroeter's SKH test [see (30)] where the bounds and beta-approximate critical points are from King (1981, Table 2) and Harrison (1982, Table 4) respectively; Szroeter's S_N test [see (31)]; the Harrison–McCabe (HM) test [see (33)], with $T_1 = I[T/2]$. Table 5 reports rejection percentages for a nominal level of 5% and 10000 replications. The MC tests are implemented with 99 simulated samples. Based on these two experiments, we make the following observations.

The BPG, K, S_N and W tests reject the null less frequently than implied by their nominal size, particularly in small samples. The G Wald-type test and the Harrison approximate SKH test have a tendency to over-reject. The bounds tests based on the HM and SKH statistics are inconclusive in a large proportion of cases. As expected, MC tests have the correct size. In the case of the BPG criterion, Honda's size correction improves both the reliability and the power properties of the test; the superiority of the MC technique is especially notable with small samples. Recall that whereas Honda's formula is generally effective, it is based on an asymptotic approximation; the MC test is theoretically exact in finite samples. Finally, we have observed that sorting the observations or the OLS residuals by the value of $\hat{y}$ leads to equivalent MC tests.

In order to compare tests of equal size, we only discuss the power of the MC tests. We observe that the S_N and the SKH MC tests (whose performance is very similar) are most powerful, followed closely by the S_F and the HM MC test, and by the G

Table 5
Variance as a function of exogenous variables

Test		$\sigma_t^2 = \alpha_0 + \alpha_1 x_{2t}$				$\sigma_t^2 = x_t'\alpha$			
		$T = 50$		$T = 100$		$T = 50$		$T = 100$	
		H_0	H_A	H_0	H_A	H_0	H_A	H_0	H_A
GQ	*F*	4.68	81.41	4.95	98.25	5.24	11.56	4.95	22.90
BPG	ASY	4.14	80.57	4.59	98.75	4.38	8.69	5.01	16.57
	BRT	4.64	81.67	4.71	98.79	4.19	8.39	4.95	16.22
	MC	4.99	80.86	4.58	98.36	4.74	9.54	5.02	16.30
K	ASY	4.74	75.14	4.51	97.52	4.02	7.06	4.32	13.77
	MC	4.98	74.70	4.46	96.77	5.08	8.10	4.73	14.63
W	ASY	2.60	20.20	4.42	34.64	2.60	3.45	4.42	7.53
	MC	4.67	26.70	4.65	33.99	4.67	5.98	4.65	7.93
G	ASY_F	5.09	80.04	4.66	98.82	5.46	9.04	5.14	15.21
	ASY_W	5.76	81.30	5.03	98.90	7.66	12.05	5.98	17.42
	MC	5.12	78.48	4.58	98.44	5.11	8.21	5.02	14.54
RB	ASY	5.50	80.06	5.22	97.96	5.75	11.99	5.41	21.01
	MC	4.58	77.03	4.76	97.49	5.03	11.02	4.90	19.78
S_F	MC	4.77	88.71	4.88	99.12	5.18	19.29	5.3	33.61
S_N	ASY	4.28	91.94	4.87	96.63	5.32	21.91	4.91	39.64
	MC	5.08	**92.09**	4.69	**99.51**	5.07	**21.38**	4.84	**38.90**
SKH	Beta	6.41	94.71	8.32	99.83	7.96	28.07	8.28	48.09
	Bound	0.71	74.74	1.54	98.09	0.96	6.17	1.50	21.02
	Bound inconc.	19.6	24.23	12.38	1.83	20.89	48.02	12.51	39.41
	MC	4.98	91.68	4.79	99.43	4.97	21.04	5.01	37.49
HM	Bound	0.79	61.31	1.91	94.61	0.74	4.18	1.85	16.25
	Bound inconc.	13.48	33.75	9.67	4.52	14.67	34.61	10.06	30.44
	MC	4.78	84.64	5.20	97.38	5.02	17.82	5.48	29.25

and *BPG* MC tests. The *GQ* and *RB* MC tests rank next whereas the *W* test performs very poorly. Note that the Szroeter *GQ*-type test S_F performs much better than the standard *GQ*; this is expected since the latter is based on residuals from a single regression on the whole sample. Overall, the most noticeable fact is the superiority of the Szroeter MC tests when compared to the commonly used procedures (e.g. the Breusch–Pagan TR^2 type tests).[17] As mentioned earlier, the Szroeter tests as initially proposed have not gained popularity due to their non-standard null distributions. Given the ease with which exact MC versions of these tests can be computed, this experiment clearly demonstrates that a sizable improvement in power results from replacing the commonly used LM-type tests with either Szroeter-type MC test.

[17] Similar conclusions are reported in Griffiths and Surekha (1986) with respect to S_N, the member of the Szroeter family whose null distribution is asymptotically normal. However, these authors also document the asymptotic tests' incorrect finite sample size.

5.3. Grouped heteroskedasticity

To illustrate the performance of MC tests for grouped heteroskedasticity (H_3), we follow the design of Binkley (1992). The model used is (1) with $T_i = 15,\ 25,\ 50;\ k_i = 4,\ 6,\ 8;\ m = 4$.[18] The regressors were drawn [only once] from a $U(0, 10)$ distribution and differed across subgroups.[19] The regression coefficients were set to one, and the variances across groups were selected so that $\delta = \sigma^2_{\max}/\sigma^2_{\min} = 1, 3, 5$, with the intermediate variances set at equal intervals, where $\sigma^2_{\min}$ and $\sigma^2_{\max}$ represent respectively the smallest and largest error variance among the m groups. The errors were drawn from the normal distribution. We considered the LR statistic, the Breusch–Pagan and Koenker statistics, the Cochran and Hartley criteria (C, H, C_r, and H_r). We also studied alternative likelihood-based test criteria introduced in Binkley (1992, p. 565), namely $LR1$, $LR2$, $LR3$ and BPG_2,[20] and considered as well a Koenker-type adjustment to BPG_2 (which we denote K_u). Empirical rejections for a nominal size of 5% in 10000 replications are summarized in Table 6.[21] The MC tests are obtained with 99 simulated samples.

Our findings can be summarized as follows. In general, LM-type asymptotic tests are undersized, whereas the asymptotic LR-type tests tend to over-reject. The variants of the LM and LR tests based on residuals from individual regressions are over-sized. As expected, size problems are more severe with small samples. The behavior of the size-corrected BPG appears to be satisfactory. Note however that we have applied the latter modification technique to BPG_u and verified that it still yields over-rejections. Indeed, we observed empirical type I errors of 12.75, 11.14 and 8.37 for $T_i = 15, 25$ and 50, respectively. Finally, the empirical size of the Cochran and Hartley statistics exceeds the nominal size. In contrast, the MC versions of all the tests considered achieve perfect size control.

In order to compare tests of equal size, we again only discuss the power of the MC tests. First of all, we observe that the MC technique improves the effective power of the LM tests. Although the correction from Honda (1988) achieves a comparable effect, its application is restricted to the standard BPG criterion. Secondly, comparing the LR and QLR tests, there is apparently no advantage to using full maximum likelihood estimation [for a similar observation in the context of SURE models, see Dufour and Khalaf (2002a, 2003)]. In general, the tests may be ranked in terms of power as follows. LR, QLR, BPG and H performed best, followed quite closely by the K and C.

[18] Results with $m = 2$ and $m = 4$ are presented in Dufour et al. (2001), where we also study the Glejser and White tests, and the Goldfeld–Quandt test when $m = 2$.

[19] We considered other choices for the design matrices, including Cauchy, lognormal, and identical regressors (across subgroups) and obtained qualitatively similar results.

[20] LR1 is obtained as in (27) replacing s_i^2 by estimates of group variances from partitioning s^2. $LR2$ is obtained as in (27) replacing s_i^2 by variance estimates from separate regressions, over the sample subgroups, and s^2 by a weighted average of these. $LR3$ is obtained like $LR2$, using unbiased variance estimates. BPG_2 is a variant of the BPG test for H_3 based on residuals from individual group regressions.

[21] For convenience, our notation differs from Binkley (1992). The QLR test refers to Binkley's $LR1$, the LR_u ($ASY1$) and LR_u ($ASY2$) refer to $LR2$ and $LR3$ tests; BPG_u corresponds to BPG_2. Note that $LR3$ obtains as a monotonic transformation of $LR2$, which yields the same MC test.

Table 6
Grouped heteroskedasticity ($m = 4$)

δ	T_i	*LR*		*QLR*		LR_u			C_r	H_r	*C*	
		ASY	MC	ASY	MC	ASY$_1$	ASY$_2$	MC	MC	MC	Tab.	MC
1	15	9.70	5.25	5.70	5.30	14.41	5.75	5.34	5.60	5.30	12.52	5.37
	25	8.05	4.64	5.23	4.92	11.82	5.35	5.01	4.96	4.94	7.25	4.97
	50	6.70	5.05	5.24	4.90	9.02	4.98	4.93	4.65	5.26	4.34	4.92
3	15	46.38	**34.23**	36.42	34.04	49.70	31.98	29.03	24.83	32.19	37.88	22.28
	25	67.58	**57.20**	59.13	56.94	68.72	52.91	50.93	39.34	55.89	42.43	34.14
	50	94.03	**91.31**	92.09	91.10	93.27	89.02	87.82	68.41	90.42	64.77	65.01
5	15	76.88	**65.17**	66.65	63.34	77.78	60.85	56.69	38.98	63.37	52.99	34.99
	25	94.59	**90.65**	91.22	89.81	94.18	87.65	85.47	60.91	89.96	63.55	54.60
	50	99.99	**99.95**	99.98	99.93	99.98	99.89	99.86	91.13	99.96	89.09	88.34
δ	T_i	*BPG*			BPG_u		*K*		K_u		*H*	
		ASY	BRT	MC	ASY	MC	ASY	MC	ASY	MC	Tab.	MC
1	15	4.52	5.47	5.44	11.06	5.30	4.41	5.47	10.37	5.31	12.97	5.04
	25	4.51	4.99	5.08	9.97	4.82	4.19	4.77	9.77	4.90	9.77	5.20
	50	4.48	4.83	4.91	8.02	4.63	4.38	4.48	7.82	4.94	3.41	5.26
3	15	28.37	32.16	31.34	41.13	26.92	22.78	25.49	34.47	21.96	46.20	27.63
	25	52.41	54.74	53.27	61.87	45.54	45.09	46.56	55.90	40.47	63.92	48.70
	50	89.53	90.07	88.91	91.11	85.59	86.55	85.67	89.08	81.66	90.58	87.30
5	15	52.12	56.23	54.70	65.42	47.03	40.12	43.79	53.63	36.31	75.51	56.81
	25	84.81	86.17	84.58	89.37	77.60	75.14	76.00	82.62	67.70	93.13	85.91
	50	99.88	99.91	99.82	99.88	99.59	99.53	99.27	99.68	99.80	99.95	99.82

Overall, no test is uniformly dominated. The MC tests constructed using variance estimates from separate regressions have a slight power disadvantage. This is somewhat expected, since the simulated samples where drawn imposing equality of the individual regression coefficients. Finally, note that the MC Hartley's test compared favorably with the LM and LR test. This, together with the fact that it is computationally so simple, suggest that applying the MC technique to Hartley's criterion yields a very useful test.

5.4. Tests for break in variance

The model used is (1) with: $T = 50$ and $k = 6$.[22] The following specification for the error variance was considered: $\sigma_t^2 = \sigma_1$, if $t \leqslant \tau_0$, and $\sigma_t^2 = \sigma_1 + \delta$, if $t > \tau_0$, where $\delta \geqslant 0$ and τ_0 is the break time (assumed unknown). The regressors and the regression coefficients parameters were chosen as in Section 5.3. Furthermore $\alpha_0 = 1$, and δ and τ_0 were set so that: $(\sigma_1 + \delta)/\sigma_1 = 1,\ 4,\ 16$, and $\tau_0/T = 0.3,\ 0.5,\ 0.7$. We

[22] Results with $T = 25$ and $T = 100$ are presented in Dufour et al. (2001), where we also study the K, W, G, RB, and HM tests. We have observed that the power of several tests converged to one with $T = 100$; we have thus chosen to report the results with $T = 50$ to allow useful power comparisons.

Table 7
Break in variance at unknown points

$T=50$		H_0	$\tau_0/T=0.3$		$\tau_0/T=0.5$		$\tau_0/T=0.7$	
σ_2/σ_1		1	2	4	2	4	2	4
GQ	F	5.5	35.0	59.9	71.2	99.8	57.1	99.3
BPG	ASY	5.3	32.6	59.3	67.0	96.5	73.7	99.6
	BRT	5.5	33.7	60.6	68.5	96.7	75.0	99.7
	MC	5.8	33.7	59.1	67.2	96.3	73.2	99.6
S_F	MC	7.2	47.0	70.3	**82.9**	99.5	73.3	99.4
S_N	ASY	5.6	47.7	72.3	80.8	99.1	82.3	99.9
	MC	6.2	45.5	70.8	77.4	98.1	80.4	99.7
SKH	Beta	9.0	58.5	79.5	87.9	99.8	87.1	1.0
	Bound	1.3	18.8	41.3	53.9	94.3	59.4	99.3
	Bound inconc.	22.8	64.5	52.0	42.7	5.7	36.0	0.7
	MC	6.8	46.7	71.9	81.9	99.0	81.1	99.8
Tests maximized over the whole sample								
$F_\times(BPG;\hat{J}^S_{(4)})$	MC	5.3	16.1	26.1	45.3	86.7	72.5	99.5
$F_{\min}(BPG;\hat{J}^S_{(4)})$	MC	5.5	12.7	18.0	31.7	66.5	62.6	98.7
$F_\times(GQ;\hat{J}^S_{(4)})$	MC	5.6	56.8	**98.6**	79.1	**100**	73.3	99.7
$F_{\min}(GQ;\hat{J}^S_{(4)})$	MC	6.0	50.0	98.2	71.6	99.8	67.2	99.4
Tests maximized over a sub-sample								
$F_\times(BPG;\hat{J}^S_{(4)})$	MC	6.0	38.0	79.8	76.9	99.4	**81.7**	**99.9**
$F_{\min}(BPG;\hat{J}^S_{(4)})$	MC	5.8	37.1	77.8	75.7	99.2	79.4	99.9
$F_\times(GQ;\hat{J}^S_{(4)})$	MC	5.4	**60.7**	98.0	80.0	99.9	74.9	99.7
$F_{\min}(GQ;\hat{J}^S_{(4)}$	MC	6.2	53.1	98.3	78.7	**100**	75.0	99.5

applied the MC versions of the standard tests GQ and BPG (using artificial regressions on $z_t = t, 1 \leqslant t \leqslant T$), S_F, SKH, and S_N tests, as well as the proposed combined tests $F_{\min}(GQ;K)$, $F_{\min}(BPG;J)$, $F_\times(BPG;\hat{J}_{(m)})$, $F_\times(GQ;\hat{K}_{(m)})$ with $m=4$. For each one of the combined tests, we considered two possible "windows" (J,K). The first one is a relatively uninformative "wide" window: $J^A=\{1,\ldots,T-1\}$, $K^A = S_1(T,T_2,k+1,T-T_2-k-1)$, with $T_2=[T/5]$. The second set of windows were based on a predetermined interval around the true break-date, namely we considered: $J^S=\{L_0,\ L_0+1,\ldots,U_0\}$, $K^S=S_1(T,T_2,\tau_0^L(k),\tau_0^L(k))$, where $T_2=[T/5]$, $\tau_0^L(k)=\max\{k+1,\tau_0-I[T/5]\}$, $\tau_0^U(k)=\min\{T-k-T_2,\tau_0+I[T/5]\}$. This yields the statistics $F_\times(BPG;\hat{J}^i_{(4)})$, $F_{\min}(BPG;J^i)$, $F_\times(GQ;\hat{K}^i_{(4)})$, $F_{\min}(GQ;K^i)$, $i=A,S$. The results are reported in Table 7.

As expected, the MC versions of all the tests achieve perfect size control. The results on relative power across tests agree roughly with those from the other experiments. Two points are worth noting. First, a remarkable finding here is the good performance of the Szroeter-type MC tests, which outperform commonly used tests such as the BPG and the GQ tests. For $\tau_0/T=0.3$, the S_F test has the best power. Second, the combined criteria perform well, and in several cases exhibit the best performance. Among these tests, product-type combined criteria perform better than min-type. The combined GQ criteria clearly dominate the standard GQ; the same holds true for the BPG-based tests, if the search window is not uninformative. Power increases substantially, where we consider the sup-tests maximized over the shorter, more informative window. These results have much to recommend the intuitively appealing combined tests, in association with the MC test method, in order to deal with problems of unknown shift in variance.

6. Conclusion

In this paper, we have described how finite-sample homoskedasticity tests can be obtained for a regression model with a specified error distribution. The latter exploit the MC test procedure which yields simulation-based exact randomized p-values irrespective of the number of replications used. The tests considered include tests for GARCH-type heteroskedasticity and sup-type tests against breaks in variance at unknown points. On observing that all test criteria are pivotal, the problem of "robustness to estimation effects" emphasized in Godfrey (1996) becomes irrelevant from our viewpoint. It is important to note that the general approach used here to obtain exact tests is not limited to the particular case of normal errors. In particular, the method proposed allows one to consider non-normal—possibly heavy-tailed (e.g., Cauchy)—error distributions, for which standard asymptotic theory would not apply.

The results of our simulation experiments suggest that Hartley-type and Szroeter-type tests seem to be the best choice in terms of power. Such tests have not gained popularity given the non-standard null distribution problem which we have solved here. We have introduced various MC combined tests, based on the minimum (sup-type tests) or the product (Fisher's combination method) of a set of p-values, and demonstrated their good performance. Although the particular test statistics considered here are designed against a two-regime variance, it would be straightforward to implement, with similar MC methods, statistics aimed at detecting a larger number of variance regimes. Finally, in the context of conditional heteroskedasticity, we have solved the unidentified nuisance parameter problem relating to ARCH-M testing.

The test procedures presented in this paper are provably valid (in finite samples) for parametric regression models with fixed (or stochastic strictly exogenous) regressors as described in assumptions (1)–(4). To the extent that the test statistics considered have the same asymptotic distribution under less restrictive regularity conditions, it is straightforward to see that the simulation-based tests presented here will also be asymptotically valid under these assumptions [for further discussion of this general asymptotic validity, see Dufour and Kiviet (1998, 2002)]. It would undoubtedly be of interest to develop similar finite-sample procedures that would be applicable to other models of econometric interest, such as: (1) dynamic models; (2) models with endogenous explanatory variables (simultaneous equations); (3) nonparametric models (especially with respect to the assumptions made on the disturbance distribution); (4) nonlinear models. These setups go beyond the scope of the present paper and are the topics of ongoing research.

Acknowledgements

The authors thank Marie-Claude Beaulieu, Bryan Campbell, Judith Giles, Jeong-Ryeol Kim and Victoria Zinde-Walsh, two anonymous referees, an Associate Editor and the Editor Ronald Gallant for several useful comments, and Jean-François Bilodeau for research assistance. This work was supported by the Canada Research Chair Program (Chair in Econometrics, Université de Montréal), the Alexander-von-Humboldt Foundation (Germany), the Institut de Finance mathématique de Montréal (IFM2), the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Canada Council for the Arts (Killam Fellowship), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, the Fonds de recherche sur la société et la culture (Québec), and the Fonds de recherche sur la nature et les technologies (Québec). This paper was also partly written at the Centre de recherche en Économie et Statistique (INSEE, Paris), the Technische Universität Dresden (Fakultät Wirtschaftswissenschaften) and the Deutsche Bundesbank (Frankfurt).

References

Ali, M.M., Giaccotto, C., 1984. A study of several new and existing tests for heteroscedasticity in the general linear model. Journal of Econometrics 26, 355–373.

Andrews, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821–856.

Andrews, D.W.K., 2001. Testing when a parameter is on the boundary of the maintained hypothesis. Econometrica 69, 683–733.

Andrews, D.W.K., Ploberger, W., 1995. Admissibility of the likelihood ratio test when a parameter is present only under the alternative. The Annals of Statistics 23, 1383–1414.

Barnard, G.A., 1963. Comment on ‘the spectral analysis of point processes’ by M.S. Bartlett. Journal of the Royal Statistical Society, Series B 25, 294.

Bartlett, M.S., 1937. Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London A 160, 268–282.

Beg, R., Silvapulle, M., Silvapulle, P., 2001. Tests against inequality constraints when some nuisance parameters are present only under the alternative: test of ARCH in ARCH-M models. Journal of Business and Economic Statistics 19, 245–253.

Bera, A.K., Ra, S., 1995. A test for the presence of conditional heteroscedasticity within ARCH-M framework. Econometric Reviews 14, 473–485.

Bewley, R., Theil, H., 1987. Monte Carlo testing for heteroscedasticity in equation systems. Advances in Econometrics 6, 1–15.

Bickel, P.J., 1978. Using residuals robustly I: tests for heteroscedasticity, non-linearity. The Annals of Statistics 6, 266–291.

Binkley, J.K., 1992. Finite sample behaviour of tests for grouped heteroscedasticity. The Review of Economics and Statistics 74, 563–568.

Bollerslev, T., Engle, R.F., Nelson, D.B., 1994. ARCH models. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, Vol. IV. North-Holland, Amsterdam, pp. 2959–3038 (Chapter 49).

Breusch, T.S., Pagan, A.R., 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica 47, 1287–1294.

Buse, A., 1984. Tests for additive heteroscedasticity: Goldfeld and Quandt revisited. Empirical Economics 9, 199–216.

Christiano, L.J., 1992. Searching for a break in GNP. Journal of Business and Economic Statistics 10, 237–249.

Cochran, W.G., 1941. The distribution of the largest of a set of estimated variances as a fraction of their total. Annals of Eugenics 11, 351–361.
Cribari-Neto, F., Ferrari, S.L.P., 1995. Bartlett-corrected tests for heteroskedastic linear models. Economics Letters 48, 113–118.
Cribari-Neto, F., Zarkos, S.G., 1999. Bootstrap methods for heteroscedastic regression models: evidence on estimation and testing. Econometric Reviews 18 (2), 211–228.
Dagenais, M.G., Dufour, J.-M., 1991. Invariance, nonlinear models and asymptotic tests. Econometrica 59, 1601–1615.
Davidson, R., MacKinnon, J.G., 1993. Estimation and Inference in Econometrics. Oxford University Press, New York.
Davies, R.B., 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64, 247–254.
Davies, R.B., 1987. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 74, 33–43.
Davison, A., Hinkley, D., 1997. Bootstrap Methods and Their Application. Cambridge University Press, Cambridge, UK.
Demos, A., Sentana, E., 1998. Testing for GARCH effects: a one-sided approach. Journal of Econometrics 86, 97–127.
Dufour, J.-M., 1989. Nonlinear hypotheses, inequality restrictions, and non-nested hypotheses: exact simultaneous tests in linear regressions. Econometrica 57, 335–355.
Dufour, J.-M., 1990. Exact tests and confidence sets in linear regressions with autocorrelated errors. Econometrica 58, 475–494.
Dufour, J.-M., Dagenais, M.G., 1992. Nonlinear models, rescaling and test invariance. Journal of Statistical Planning and Inference 32, 111–135.
Dufour, J.-M., Khalaf, L., 2001. Monte Carlo test methods in econometrics. In: Baltagi, B. (Ed.), Companion to Theoretical Econometrics Blackwell Companions to Contemporary Economics. Basil Blackwell, Oxford, UK, pp. 494–519 (Chapter 23).
Dufour, J.-M., Khalaf, L., 2002a. Exact tests for contemporaneous correlation of disturbances in seemingly unrelated regressions. Journal of Econometrics 106 (1), 143–170.
Dufour, J.-M., Khalaf, L., 2002b. Simulation based finite and large sample tests in multivariate regressions. Journal of Econometrics 111 (2), 303–322.
Dufour, J.-M., Khalaf, L., 2003. Finite sample tests in seemingly unrelated regressions. In: Giles, D.E.A. (Ed.), Computer-Aided Econometrics. Marcel Dekker, New York, pp. 11–35 (Chapter 2).
Dufour, J.-M., Kiviet, J.F., 1996. Exact tests for structural change in first-order dynamic models. Journal of Econometrics 70, 39–68.
Dufour, J.-M., Kiviet, J.F., 1998. Exact inference methods for first-order autoregressive distributed lag models. Econometrica 66, 79–104.
Dufour, J.-M., Kiviet, J.F., 2002. Generic asymptotic validity of Monte Carlo tests. Technical Report, CIRANO and CIREQ, Université de Montréal, and Tinbergen Institute, University of Amsterdam.
Dufour, J.-M., Torrès, O., 1998. Union-intersection and sample-split methods in econometrics with applications to SURE and MA models. In: Giles, D.E.A., Ullah, A. (Eds.), Handbook of Applied Economic Statistics. Marcel Dekker, New York, pp. 465–505.
Dufour, J.-M., Torrès, O., 2000. Markovian processes, two-sided autoregressions and exact inference for stationary and nonstationary autoregressive processes. Journal of Econometrics 99, 255–289.
Dufour, J.-M., Khalaf, L., Bernard, J.-T., Genest, I., 2001. Simulation-based finite-sample tests for heteroskedasticity and ARCH effects. Technical Report, C.R.D.E., Université de Montréal, and GREEN, Université Laval. 48pp.
Durbin, J., Watson, G.S., 1950. Testing for serial correlation in least squares regression I. Biometrika 37, 409–428.
Dwass, M., 1957. Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics 28, 181–187.
Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50 (4), 987–1008.

Engle, R.F. (Ed.), 1995. ARCH: Selected Readings Advanced Texts in Econometrics. Oxford University Press, Oxford, U.K.

Engle, R.F., Hendry, D.F., Trumble, D., 1985. Small-sample properties of ARCH estimators and test. Canadian Journal of Economics 18, 66–93.

Evans, M.A., 1992. Robustness of size of tests of autocorrelation and heteroscedasticity to nonnormality. Journal of Econometrics 51, 7–24.

Evans, M.A., King, M.L., 1985a. Critical value approximations for tests of linear regression disturbances. Australian Journal of Statistics 27, 68–83.

Evans, M.A., King, M.L., 1985b. A point optimal test for heteroscedastic disturbances. Journal of Econometrics 27, 163–178.

Farebrother, R.W., 1987. The statistical foundation of a class of parametric tests for heteroscedasticity. Journal of Econometrics 36, 359–368.

Fisher, R.A., 1932. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

Folks, J.L., 1984. Combination of independent tests. In: Krishnaiah, P.R., Sen, P.K. (Eds.), Handbook of Statistics 4: Nonparametric Methods. North-Holland, Amsterdam, pp. 113–121.

Giaccotto, C., Sharma, S.C., 1988. Jackknife tests for heteroscedasticity in the general linear model. Australian Journal of Statistics 30, 200–216.

Glejser, H., 1969. A new test for heteroscedasticity. Journal of the American Statistical Association 64, 316–323.

Godfrey, L.G., 1978. Testing for multiplicative heteroscedasticity. Journal of Econometrics 8, 227–236.

Godfrey, L.G., 1988. Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches. Cambridge University Press, Cambridge, UK.

Godfrey, L.G., 1996. Some results on the Glejser and Koenker tests of heteroscedasticity. Journal of Econometrics 72, 275–299.

Goldfeld, S.M., Quandt, R., 1965. Some tests for heteroscedasticity. Journal of the American Statistical Association 60, 539–547.

Gouriéroux, C., 1997. ARCH Models and Financial Applications Springer Series in Statistics. Springer, New York.

Griffiths, W.E., Surekha, K., 1986. A Monte Carlo evaluation of the power of some tests for heteroscedasticity. Journal of Econometrics 31, 219–231.

Hansen, B.E., 1996. Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64, 413–430.

Hansen, B.E., 2000. Testing for structural change in conditional models. Journal of Econometrics 97 (1), 93–115.

Harris, P., 1985. An asymptotic expansion for the null distribution of the efficient score statistic. Biometrika 72, 653–659.

Harrison, M.J., 1980. The small sample performance of the Szroeter bounds test for heteroscedasticity and a simple test for use when Szroeter's test is inconclusive. Oxford Bulletin of Economics and Statistics 42, 235–250.

Harrison, M.J., 1981. A comparison of the bounds, beta-approximate, and exact variants of two tests for heteroscedasticity based on ordinary least squares residuals. The Economic and Social Review 12, 235–252.

Harrison, M.J., 1982. Table of critical values for a beta approximation to Szroeter's statistic for testing heteroscedasticity. Oxford Bulletin of Economics and Statistics 44, 159–166.

Harrison, M.J., McCabe, B.P.M., 1979. A test for heteroscedasticity based on ordinary least squares residuals. Journal of the American Statistical Association 74, 494–499.

Hartley, H.O., 1950. The maximum F-ratio a short-cut test for heterogeneity of variances. Biometrika 37, 308–312.

Harvey, A.C., 1976. Estimating regression models with multiplicative heteroscedasticity. Econometrica 44, 461–465.

Harvey, A.C., Phillips, G.D.A., 1974. A comparison of the power of some tests for heteroscedasticity in the general linear model. Journal of Econometrics 2, 307–316.

Honda, Y., 1988. A size correction to the Lagrange multiplier test for heteroscedasticity. Journal of Econometrics 38, 375–386.

Hong, Y., Shehadeh, R.D., 1999. A new test for ARCH effects and its finite-sample performance. Journal of Business and Economic Statistics 17 (1), 91–108.

Imhof, P.J., 1961. Computing the distribution of quadratic forms in normal variables. Biometrika 48, 419–426.

Jöckel, K.-H., 1986. Finite sample properties and asymptotic efficiency of Monte Carlo tests. The Annals of Statistics 14, 336–347.

King, M.L., 1981. A note on Szroeter's bound test. Oxford Bulletin of Economics and Statistics 43, 315–326.

Kiviet, J.F., Dufour, J.-M., 1997. Exact tests in single equation autoregressive distributed lag models. Journal of Econometrics 80, 325–353.

Koenker, R., 1981. A note on studentizing a test for heteroscedasticity. Journal of Econometrics 17, 107–112.

Koenker, R., Bassett, G., 1982. Robust tests for heteroscedasticity based on regression quantiles. Econometrica 50, 43–61.

Lee, J.H.H., 1991. A Lagrange multiplier test for GARCH models. Economics Letters 37, 265–271.

Lee, J.H., King, M.L., 1993. A locally most mean powerful based score test for ARCH and GARCH regression disturbances. Journal of Business and Economic Statistics 11, 17–27 (Correction 12 (1994) 139).

LeRoy, S.F., 1996. Stock price volatility. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics 14: Statistical Methods in Finance. North-Holland, Amsterdam, pp. 193–208.

Lumsdaine, R., 1995. Finite-sample properties of the maximum likelihood estimator in GARCH(1,1) and IGARCH(1,1) models: a Monte Carlo investigation. Journal of Business and Economic Statistics 13, 1–10.

MacKinnon, J.G., White, H., 1985. Some heteroscedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics 29, 305–325.

Maekawa, K., 1988. Comparing the Wald, LR and LM tests for heteroscedasticity in a linear regression model. Economic Letters 26, 37–41.

McCabe, B.P., 1986. Testing for heteroscedasticity occurring at unknown points. Statistics-Theory and Methods 15, 1597–1613.

Newey, W.K., Powell, J.L., 1987. Asymmetric least squares estimation and testing. Econometrica 55, 819–847.

Pagan, A.R., Hall, A.D., 1983. Diagnostic tests as residual analysis. Econometric Reviews 2, 159–218 (Comments and Reply, 219–254).

Pagan, A.R., Pak, Y., 1993. Testing for heteroskedasticity. In: Maddala, G.S., Rao, C.R., Vinod, H.D. (Eds.), Handbook of Statistics 11: Econometrics. North-Holland, Amsterdam, pp. 489–518.

Palm, F.C., 1996. GARCH models of volatility. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics 14: Statistical Methods in Finance. North Holland, Amsterdam, pp. 209–240.

Pearson, E.S., Hartley, H.O., 1976. Biometrika Tables for Statisticians, 3rd Edition. Biometrika Trust, London, UK.

Pearson, K., 1933. On a method of determining whether a sample of size n supposed to have been drawn from a parent population. Biometrika 25, 379–410.

Ramsey, J.B., 1969. Tests for specification errors in the classical linear least squares regression analysis. Journal of the Royal Statistical Society Series B 31, 350–351.

Rivest, L.-P., 1986. Bartlett's, Cochran's, and Hartley's tests on variances are liberal when the underlying distribution is long-tailed. Journal of the American Statistical Association 81, 124–128.

Shaban, S.A., 1980. Change point problem and two-phase regression: an annotated bibliography. International Statistical Review 48, 83–93.

Sharma, S.C., Giaccotto, C., 1991. Power and robustness of jackknife and likelihood-ratio tests for grouped heteroscedasticity. Journal of Econometrics 49, 343–372.

Sullivan, M., Giles, D.E.A., 1995. The robustness of ARCH/GARCH tests to first-order autocorrelation. Journal of Quantitative Economics 11, 35–61.

Szroeter, J., 1978. A class of parametric tests for heteroscedasticity in linear econometric models. Econometrica 46, 1311–1327.

Theil, H., 1971. Principles of Econometrics. Wiley, New York.

Tippett, L.H., 1931. The Methods of Statistics. Williams and Norgate, London.
Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, New York.
White, H., 1980. A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. Econometrica 48, 817–838.
Wilkinson, B., 1951. A statistical consideration in psychological research. Psychology Bulletin 48, 156–158.