15 June 2004

Simulation-based tests for heteroskedasticity: some further results*

Abstract

As shown by Dufour, Khalaf, Brenard and Genest, exact tests for heteroskedasticity can be obtained, by using Monte Carlo (MC) techniques, if it is assumed that the true form of the error distribution under the null hypothesis is known. The corresponding nonparametric bootstrap tests are only asymptotically valid, but do not require specification of the error law. Since information about the precise form of the error distribution is not often available to applied workers, two questions merit attention. First, how robust are MC tests to incorrect assumptions about the error distribution? Second, are nonparametric bootstrap tests markedly inferior to MC tests when the latter use the correct error distribution? Results relevant to these two questions are derived using asymptotic analysis and simulation experiments. The evidence suggests that the combination of an asymptotically pivotal test statistic with a nonparametric bootstrap gives a robust and well-behaved procedure.

L.G. Godfrey
Department of Economics and Related Studies
University of York
Heslington, York YO10 5DD
United Kingdom

C.D. Orme (corresponding author)
School of Economic Studies
University of Manchester
Manchester M13 9PL
United Kingdom
tel. +44 161 275 4856
fax +44 161 275 4928
e-mail chris.orme@man.ac.uk

J.M.C. Santos Silva
ISEG/Universidade Técnica de Lisboa
R. do Quelhas 6
1200 Lisboa
Portugal
JEL classification codes: C12, C15, C2, C52.

Key Words: Bootstrap; Distributional assumptions; Exact tests; Monte Carlo tests.

---

*Santos Silva gratefully acknowledges the partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER. The authors are grateful to Professors P. Hall and M. Titterington for helpful advice.

1 Introduction

There is an extensive literature on the construction, implementation and interpretation of tests for heteroskedasticity which informs a large body of empirical work. This literature, and a survey of the use of heteroskedasticity tests in empirical studies, is usefully summarized in a recent paper by Dufour, Khalaf, Brenard and Genest (2004). As observed in that paper, most test procedures employ asymptotically valid critical values and numerous studies have compared the finite sample behaviour of these "asymptotic" tests in order to provide applied workers with some guidance on their use. The evidence is that first order asymptotic theory, in general, provides a rather poor guide to finite sample behaviour. Rather than use asymptotic critical values, Godfrey and Orme (1999) exploit Beran's (1988) results and show that a simple bootstrap procedure can, in many cases, provide a greater degree of control over significance levels than that previously afforded by standard asymptotic theory. Indeed, "bootstrap" tests are now widely acknowledged as having the potential to provide much more reliable inferences, and their use in empirical work should increase rapidly as computing costs continue to fall and appropriate routines are added to standard packages.

In general, the bootstrap tests deliver improvements in the Error in Rejection Probability (ERP), i.e. the discrepancy between actual and desired significance levels, but they remain only asymptotically valid. In contrast, Dufour *et al.* (2004) provide results which show that, through the use of Monte Carlo techniques, it is possible to eliminate this discrepancy completely. Their analysis is in the context of tests for heteroskedasticity, but the methodology is more widely applicable. It will be useful to refer to test procedures which employ these Monte Carlo techniques as MC tests (and retain Monte Carlo study to refer to sampling experiments employed to investigate the behaviour of various test criteria). The distinct advantage of the MC method is that it provides simple exact procedures for a large class of tests and enables new tests to be defined and implemented, even if they are based on statistics whose finite sample or asymptotic distributions are intractable.

The applicability of these MC tests, however, rests on the strong assumption that the null distribution of the error of the regression model is known.[1] Given this assumption, Dufour *et al.* (2004) show that the test criteria they consider are exactly pivotal, under the null. However, if the focus is on the problem of testing for heteroskedasticity, it is not clear that applied workers will want to use tools that depend in an important way upon the validity of an auxiliary assumption that specifies the general form of the error distribution: in many cases there is no precise non-sample information upon which to base this assumption.

The aims of this paper are to analyse the consequences for MC tests of making an incorrect assumption about the distribution of the error, and to compare MC tests with nonparametric bootstrap tests when the former enjoy the benefit of correct specification of the error model. It is shown that it is necessary to distinguish between two scenarios when considering the effects of incorrect specification of the error distribution. First, the test criterion is an asymptotic pivot: it is shown below that, in this case, the MC test is only asymptotically valid (with theory predicting that it possesses an ERP which is of the same order in T, the sample size, as that of the asymptotic test). Second, the test criterion is not an asymptotic pivot: it is shown below that, in this case, the MC test is asymptotically invalid (with an ERP which is $O(1)$). In both cases the bootstrap test remains asymptotically valid. Furthermore, in the first scenario the bootstrap test has an ERP which is of smaller order in T than both the MC and asymptotic test whilst in the second scenario it has an ERP which is of the same order in T as the asymptotic test. The term "asymptotic pivot" is used in the precise sense of Beran (1988), and is defined in Section 3.

The paper is organized as follows. Section 2 defines the model and discusses the asymptotic test procedures that are employed in the analysis. In order to develop the principal results, which are generally applicable, only a subset of the tests considered by Dufour *et al.* (2004) need be considered and the focus here is on the important class of tests in which the

[1] Under this assumption, the MC method is, in fact, an example of Case 1 detailed by Horowitz (1994) when discussing the information matrix test: given regressor values, simulation techniques can be used to obtain exact inferences.

alternative hypothesis has the variances depending on exogenous variables. Section 3 describes the bootstrap and MC tests, collectively termed simulation-based tests, and provides an analysis of their asymptotic properties. In order to illustrate the theoretical findings, a Monte Carlo study is undertaken and this is described in Section 4, with the results reported in Section 5. Finally, Section 6 concludes.

2 Models and tests

2.1 Models

As in Dufour *et al.* (2004), the regression model with heteroskedastic errors is written as

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

with

$$u_t = \sigma_t \varepsilon_t, \tag{2}$$

in which $x_t = (x_{t1}, \ldots, x_{tk})'$, $x_{t1} = 1$, $\beta = (\beta_1, \ldots, \beta_k)'$, $0 < \sigma_t < \infty$, and the terms ε_t are independently and identically distributed (iid) with common cdf $\mathcal{F}$ having zero mean and unit variance, $t = 1, \ldots, T$. The null hypothesis is $H_0 : \sigma_t = \sigma$, $0 < \sigma < \infty$, for all t.

It is assumed here that, at a minimum, the regularity conditions given by Koenker (1981) are satisfied. These conditions include conventional requirements about the limiting behaviour of the regressors of (1) which are taken to be strictly exogenous; see (A.1) of Koenker (1981, p. 108). Regarding the iid random terms ε_t of (2), it is further assumed that $E(\varepsilon_t^4) = \mu_4 < \infty$, for $t = 1, \ldots, T$. The assumptions for ε_t correspond to (A.2) of Koenker (1981), except that he uses a slightly different parametrization with the ε_t being iid$(0, \sigma^2)$; so that, in his framework, the null hypothesis is $H_0^K : \sigma_t = 1$ for all t.

The results of OLS estimation of (1) play an important role in the construction of tests of H_0. Let the OLS estimator of β in (1) be

$$\hat{\beta} = \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t y_t, \tag{3}$$

with associated predicted values and residuals given by

$$\hat{y}_t = x_t'\hat{\beta} \text{ and } \hat{u}_t = y_t - \hat{y}_t,\ t = 1, \ldots, T.$$

The variance estimate for the model of H_0 is denoted by $\hat{\sigma}^2 = T^{-1}\sum_{t=1}^{T}\hat{u}_t^2$.

2.2 The test statistics

As discussed below, when assessing the consistency of H_0 with the sample data, differences in the forms of test statistics reflect differences in the assumptions made about the specification of the alternative hypothesis. When H_0 is true, tests that use different alternatives will all be (at least) asymptotically valid, provided required regularity conditions are satisfied.

In general, some tests are intended to have finite sample validity, others are designed to be only asymptotically valid: the assumptions for the former type are usually more restrictive than those for the latter. When tests are designed to be exact, employing critical values from standard statistical tables (e.g. for t or F distributions), the ε_t are typically assumed to be normally distributed; so that, in such cases, the ε_t are NID$(0, 1)$. However, as made clear by Dufour *et al.* (2004), exact tests can be derived for other distributions if MC methods, rather than standard tables, are adopted to assess statistical significance. There are consequences, however, if an incorrect assumption about the error distribution is made.

In order to develop this argument, the basic forms of the test statistics to be considered in this paper are designed for heteroskedastic alternatives in which the variances are functions of exogenous variables and share the following characteristics: (i) they do not depend upon the nuisance parameters of the null model, viz. (β', σ^2); and (ii) they are calculated using results from OLS estimation.

2.2.1 Breusch-Pagan statistic

Breusch and Pagan (1979), hereafter BP, consider an alternative hypothesis that can be written as

$$\sigma_t^2 = h(z_t'\gamma), \tag{4}$$

in which $h(\cdot)$ is a function with first and second derivatives, $z_t = (z_{t1}, \ldots, z_{tm})'$, $z_{t1} = 1$ and $\gamma = (\gamma_1, \ldots, \gamma_m)'$, $t = 1, \ldots, T$. The exogenous variables of z_t satisfy regularity conditions similar to those for the regressors of x_t in (1); see Koenker (1981, p. 108). Homoskedasticity is implied by the $(m-1)$ restrictions of $\gamma_2 = \ldots = \gamma_m = 0$. By assuming that under the null hypothesis the errors u_t are NID$(0, \sigma^2)$, BP derive a Lagrange multiplier (LM) test and show that it is based upon a test statistic equal to one half of the explained sum of squares from the OLS regression of $\hat{\sigma}^{-2}\hat{u}_t^2$ on z_t. Under this assumption, the BP test statistic is asymptotically distributed as $\chi^2(m-1)$, with significantly large values indicating that the null hypothesis is inconsistent with the sample data. This form of a LM test has been derived independently by other researchers; see Cook and Weisberg (1983), and Godfrey (1979).

The assumption of normality is important for the asymptotic validity of the BP test.[2] Moreover if any other error distribution were assumed, the log-likelihood for the alternative model, and hence the LM test statistic for $\gamma_2 = \ldots = \gamma_m = 0$, would be different. Thus, unless normality is assumed, there is little motivation for using the procedure proposed by BP.

2.2.2 Koenker's statistic

Koenker (1981) extends the BP approach to obtain a statistic that is robust to nonnormality, provided that the terms ε_t satisfy the regularity conditions described above and, in particular, have finite fourth moment. Koenker's "Studentized" test uses, as a statistic, T times the R^2 from the OLS regression of $\hat{u}_t^2$ (or equivalently $\hat{\sigma}^{-2}\hat{u}_t^2$) on z_t of (4). The $\chi^2(m-1)$ distribution now provides an asymptotically valid basis for a test of the assumption of homoskedasticity in the presence of unspecified forms of nonnormality.

[2] More precisely, it is required that $E(u_t^4) = 3\sigma^4$ under H_0, which is true when the ε_t are NID$(0, 1)$.

2.2.3 Glejser's statistic and modified forms

In Glejser's (1969) approach, the statistic is calculated by using $|\hat{u}_t|$, rather than $\hat{u}_t^2$, as a typical observation on the dependent variable in the artificial regression on z_t.[3] Therefore, Glejser's algorithm uses, say,

$$|\hat{u}_t| = z_t'\gamma + w_t. \tag{5}$$

The restrictions to be tested in (5) are, as before, $\gamma_2 = \ldots = \gamma_m = 0$ and a standard F-test of these restrictions is assumed to be asymptotically valid.[4] However, results reported by Godfrey (1996) imply that the Glejser test requires an assumption additional to those required for Koenker's large sample test.

For Glejser tests to be asymptotically valid, the error distribution under H_0 must have $\Pr(u_t \geqslant 0) = 0.5$, for which a sufficient (but not necessary) condition is that u_t has a symmetric distribution.[5] The assumption that $\Pr(u_t \geqslant 0) = 0.5$ is restrictive but it can be relaxed by modifying the dependent variable used in the artificial regression proposed by Glejser (1969). A suitable modification is proposed independently by Machado and Santos Silva (2000) and Im (2000), hereafter MSSI. The modified Glejser test is derived by replacing $|\hat{u}_t|$ in (5) by $g(\hat{u}_t)$, where

$$g(\hat{u}_t) = \hat{u}_t \times \{\mathbf{1}(\hat{u}_t \geqslant 0) - \hat{\pi}\},$$

$\mathbf{1}(\cdot)$ is the usual indicator function, and $\hat{\pi} = T^{-1}\sum_{t=1}^{T}\mathbf{1}(\hat{u}_t \geqslant 0)$ is the sample proportion of nonnegative residuals. Observe that the Glesjer tests, and variants thereof, only require the ε_t to possess finite second moments; see Machado and Santos Silva (2000, Assumption 1, p. 199).

---

[3] Glejser's test can be interpreted as the LM test for the case of errors that have a double exponential distribution.

[4] As described in Dufour *et al.* (2004), $\gamma_2 = \ldots = \gamma_m = 0$ can also be tested by comparing T times the R^2 from the OLS estimation of (5) with right-hand-side critical values from the $\chi^2(m-1)$ distribution.

[5] Pagan and Pak (1993) remark on the relevance of conditional symmetry of the error term to the asymptotic validity of the Glejser test.

2.2.4 Szroeter's test statistic

Szroeter (1978) derives a family of test statistics by adding a strong assumption concerning the alternative hypothesis to those required under the null hypothesis. His specification of the alternative model imposes the restriction that there is a known ordering of the variances σ_t^2. Consequently, after suitable reordering of the data, which is denoted by enclosing the observation subscript in round brackets, variances satisfy $\sigma_{(t)}^2 \geqslant \sigma_{(t-1)}^2$ with at least one strong inequality when there is heteroskedasticity. This sort of ordering is not, in general, implied by the alternatives adopted by BP, Koenker, Glejser, and MSSI. Moreover, it cannot be assumed that the information that permits correct ordering of variances under the alternative is routinely available.

Given his assumption about the pattern of heteroskedasticity, Szroeter (1978) considers the use of a subsample of squared residuals when calculating test statistics. Accordingly, let A be a specified non-empty subset of $\{1, \ldots, T\}$, and $\tilde{u}_{(t)}$, $t \in A$, be the corresponding subset of ordered residuals, e.g. $\tilde{u}_{(t)} = \hat{u}_{(t)}$ gives a subsample of ordered OLS residuals. Also let $h_{(t)}$, $t \in A$, be a set of nonstochastic scalars such that $h_{(t)} \leq h_{(s)}$ if $t < s$.

The general form of Szroeter's statistic before it is "centered under the null" is given by

$$\tilde{h} = \frac{\sum_{t \in A} h_{(t)} \tilde{u}_{(t)}^2}{\sum_{t \in A} \tilde{u}_{(t)}^2}. \tag{6}$$

The specific versions of Szroeter's statistic that are used by Dufour *et al.* (2004) all use $A = \{1, \ldots, T\}$ and OLS residuals; so that they are special cases of

$$\hat{h} = \frac{\sum_{t=1}^{T} h_{(t)} \hat{u}_{(t)}^2}{\sum_{t=1}^{T} \hat{u}_{(t)}^2} = \frac{\sum_{t=1}^{T} h_t \hat{u}_t^2}{\sum_{t=1}^{T} \hat{u}_t^2}. \tag{7}$$

Szroeter (1978) proposes that, given his assumption about the pattern of variances under the alternative hypothesis, $H_0 : \sigma_t^2 = \sigma^2$ for all t should be rejected if the criterion of (7) is significantly greater than

$$\bar{h} = T^{-1} \sum_{t=1}^{T} h_{(t)} = T^{-1} \sum_{t=1}^{T} h_t. \tag{8}$$

If $\hat{\phi} = \hat{h} - \bar{h}$, it is easy to show that

$$\sqrt{T}\hat{\phi} = \frac{T^{-1/2}\sum_{t=1}^{T}(\hat{u}_t^2 - \hat{\sigma}^2)h_t}{\hat{\sigma}^2}, \tag{9}$$

in which $p\lim \hat{\sigma}^2 = \sigma^2$ under homoskedasticity. The numerator of the right-hand-side of (9) is, however, just the quasi-LM criterion for testing $\gamma_2 = 0$ in the Koenker-type artificial regression, say,

$$\hat{u}_t^2 = \gamma_1 + \gamma_2 h_t + w_t. \tag{10}$$

Using $m = 2$ and $z_{t2} = h_t$ in Koenker's approach will, therefore, produce a test statistic that is equivalent to Szroeter's criterion. A one-sided Studentized score test can be obtained by using the t-ratio for testing $\gamma_2 = 0$ against $\gamma_2 > 0$ after OLS estimation of (10).

2.2.5 Goldfeld-Quandt statistics

As in Szroeter's (1978) framework, the statistic proposed by Goldfeld and Quandt (1965), hereafter GQ, is based upon the assumption that there is a known ordering by some specified exogenous rule such that $\sigma^2_{(t)} \geqslant \sigma^2_{(t-1)}$, $t = 2, \ldots, T$. The null hypothesis is $\sigma^2_{(t)} = \sigma^2_{(t-1)}$, $t = 2, \ldots, T$, and, under the alternative hypothesis, $\sigma^2_{(t)} > \sigma^2_{(t-1)}$, for at least one value of t.

After reordering according to the rule, the sample of T observations is partitioned into three subsamples: the first and third are used for separate OLS estimations and must have more than k observations; and the second is discarded. Let T_1, T_2 and T_3 denote the numbers of observations in the three subsamples, so $T = T_1 + T_2 + T_3$, with $T_1 > k$ and $T_3 > k$. The sums of squared OLS residuals from the first T_1 and last T_3 observations are denoted by S_1 and S_3, respectively. The general form of the GQ statistic is then

$$GQ(T_1, T_3, k) = \frac{S_3/(T_3 - k)}{S_1/(T_1 - k)}. \tag{11}$$

Provided the strong auxiliary assumption of normality is made, an exact test can be obtained by treating the statistic of (11) as being distributed as $F(T_3 - k, T_1 - k)$ and using critical values from the right-hand tail of this distribution. The assumption of normality

also implies that $\hat{y}_t$ and $\hat{u}_s$ are independent for all t and s. Consequently the data can be ordered according to the values of the predicted values $\hat{y}_t$, rather than a nonrandom choice, without complicating the finite sample distribution theory for GQ tests. Dufour *et al.* (2004) point out that this result will also be applicable for any nonnormal error distribution that implies the independence of OLS predicted and residual values. Effects on the behaviour of GQ tests, based upon ordering by $\hat{y}_t$, when the errors are not normal and independence of OLS predicted and residual values cannot be guaranteed are examined below.

For future reference, it is important to investigate whether the GQ statistic in (11) is an asymptotic pivot.[6] Under the null, it is $O_p(1)$ but has a degenerate limit null cdf, since its probability limit, as $T \to \infty$, is clearly $\dfrac{\sigma^2}{\sigma^2} = 1$, on the assumption that both T_1 and T_3 are proportional to T. Therefore, in order to construct a statistic with a non-degenerate distribution, centering and norming are required. Assuming, following standard practice and Dufour *et al.* (2004), that $T_1 = T_3 \propto T$, it is straightforward to show that, writing $GQ \equiv GQ(T_1, T_3, k)$,

$$\sqrt{T_1}\,(GQ - 1) = \frac{1}{\sqrt{T_1}} \left(\frac{S_3 - S_1}{\sigma^2} \right) + o_p(1). \tag{12}$$

Under fairly standard conditions, the first term on the right hand side of (12) is $O_p(1)$ with a limit null distribution which is normal, zero mean but with a variance depending upon $\mathcal{F}$, the error distribution, through its fourth moment. Hence, the centred and normed GQ test statistic is not an asymptotic pivot. Note that: (i) this result applies for unspecified error distributions when the ordering rule is exogenous; and (ii) the statistic defined by (12) is a known linear transformation of GQ with a positive slope coefficient, so that p-values of the two statistics are equivalent and there is no need to work with the former, rather than the latter.

[6] The classification of the other test statistics discussed in this subsection as asymptotically pivotal/non-pivotal has been examined (explicitly or implicitly) by Godfrey and Orme (1999). If calculated by using a t-test of $\gamma_2 = 0$ in the Koenker-type regression (10), the Szroeter test statistic based on $\sqrt{T}\phi$ of (9) is asymptotically pivotal.

2.3 Comparisons of powers of tests

While the main purpose of this paper is to provide evidence and comments on the finite sample significance levels of tests for heteroskedasticity, a brief discussion of power comparisons is worthwhile. There are several difficulties associated with comparing powers and attempting to draw conclusions about the relevant merits of different procedures.

First, it is important that power comparisons are not contaminated by important differences in significance levels. A thoughtful discussion of empirically relevant critical values for Monte Carlo studies of power is given by Horowitz and Savin (2000). They conclude that, in general, bootstrap methods should be used to estimate critical values. This conclusion is supported, in the context of tests for heteroskedasticity, by the Monte Carlo results of Godfrey and Orme (1999). When the true error distribution is known (apart from the value of σ^2), the MC tests of Dufour *et al.* (2004) are exact and so there are no differences in significance levels to impair comparisons of their powers.

Second, apparently general remarks about, say, "the" Koenker test, "the" Szroeter test, and "the" GQ test can be misleading because each procedure requires specific choices to be made for its implementation. For BP, Glejser and Koenker tests, the variables of z_t must be specified. For Szroeter's test, an ordering must be selected, along with the single test variable h_t. For GQ-type checks, the values of T_1, T_2 and T_3 must be chosen after applying a specified rule for reordering the data y_t (or equivalently the residuals $\hat{u}_t$).

Third, there is the complicating factor that tests being compared may be based upon different alternatives, with not all (or even any) of these alternatives being (locally) equivalent to the true variance model. In order to obtain a fair basis for comparisons, attention must be given to what is being assumed about information concerning the alternative. Farebrother (1987) provides some interesting comments on the interaction between the information about the alternative that is used and the power of a test. Godfrey (1996, Appendix 1) gives re-

sults on the effects on asymptotic local power of using an incorrect alternative model that are pertinent to the comparison of tests.

There will clearly be many possible outcomes of power comparisons for different combinations of assumed and true alternative models. For example, as explained in Subsection 2.2.4, Szroeter's test can be carried out using a one-sided t-test in the context of a Koenker-type artificial regression in which the regressors are an intercept term and the selected scalar h_t. The results of Godfrey (1996, Appendix 1) imply that, under a sequence of contiguous alternatives given by

$$\sigma_t^2 = \sigma^2 + T^{-1/2}(v_t'\rho),$$

in which v_t is a p-dimensional vector of exogenous variables and $0 < \rho'\rho < \infty$, the asymptotic local power of Szroeter's test with test variable h_t can be less than, equal to, or greater than that of a Koenker check that uses the test variables of some vector z_t. There is no generally valid ranking by asymptotic local power: given v_t and ρ, the outcome of a comparison will depend upon the choices made for h_t and z_t in the Szroeter and Koenker tests, respectively. In the special case in which $p = 1$, $\rho_1 \geqslant 0$ and $v_{t1} = h_t$, there is an unambiguous result which is that the asymptotic local power of a general Koenker-type test cannot be greater that of the Szroeter test.

Given the same information about the alternative, differences between tests proposed by different authors are often of small order or reflect differences in auxiliary assumptions about distributions. For example, the version of Szroeter's statistic given by Griffiths and Surekha (1986, eq. 5) is the positive square root of a modified form of Koenker's TR^2 statistic: the modification is that the sample variance of squared residuals $\hat{u}_t^2$ is replaced by $2\hat{\sigma}^4$ which is valid when the errors of (1) are NID$(0, \sigma^2)$.

In practical situations, applied workers may know little about the true variance model (otherwise such knowledge would have been incorporated in the model's specification). They may, therefore, have to resort to general information-parsimonious rules for choosing test variables. Szroeter (1978) suggests the general purpose choice of $h_{(t)} = t$, which, in combi-

nation with the rule for ordering, implies the values of the terms h_t. However, other choices are possible; see Dufour *et al.* (2004, eqs. 30-32). Many choices of z_t in Koenker's procedure can be made (including, when it is thought useful, Szroeter's suggested variable with an intercept term). Many researchers are likely to use x_t of (1) to construct z_t. Two obvious choices of this type are $\{m = k, z_t = x_t\}$ and $\{m = r, z_t \stackrel{\circ}{=} x_t \otimes x_t\}$, where, in the latter case, r is the number of nonredundant variables in $x_t \otimes x_t$ and $\stackrel{\circ}{=}$ indicates an equality that holds after redundant variables are omitted. (The second choice for z_t gives White's (1980) general check for heteroskedasticity.)

It is worth noting that, when discussing the behaviour of tests under alternatives in which variances are functions of exogenous variables, Dufour *et al.* (2004) focus on two cases. First, variances are assumed to increase monotonically with the values of a single regressor. Second, the variances are assumed to increase monotonically with $E(y_t|x_t)$. These two special cases are obviously quite restrictive in the context of multiple regression models. Consequently the generality of their remarks on the relative powers of different tests is limited.

3 Simulation-based tests

Here simulation-based tests, and their properties, are discussed. In general, let a typical test statistic be denoted by τ. Under the null hypothesis, τ is assumed to be $O_p(1)$ with a non-degenerate limit null distribution that satisfies Beran's (1988) conditions given below. The observed value of τ is denoted $\tau_{(0)}$. When examining the asymptotic behaviour of τ, an important consideration is whether or not τ is asymptotically pivotal, as defined by Beran (1988). Specifically, if τ is an asymptotic pivot, its limit null distribution is independent of $(\beta', \sigma^2, \mathcal{F})$, within the context of (1) and (2). Whilst for all standard forms of τ discussed in this paper there is always independence with respect to (β', σ^2), for some of these test statistics the limit null distribution crucially depends upon $\mathcal{F}$. In the latter case, τ is not asymptotically pivotal.

3.1 Nonparametric bootstrap

The nonparametric bootstrap provides a simulation-based method of obtaining asymptotically valid inferences without assuming precise knowledge of $\mathcal{F}$. Here, the unknown true cdf for the errors is estimated by its empirical counterpart, $\hat{\mathcal{F}}_T$, defined as follows:

$$\hat{\mathcal{F}}_T\text{: probability } \frac{1}{T} \text{ on } \hat{u}_t,\ t = 1, \ldots, T.$$

Observe that the residuals used to define $\hat{\mathcal{F}}_T$ do not need to be centred since it is assumed that (1) contains an intercept, thus implying $\sum_t \hat{u}_t = 0$. Freedman (1981) shows that $d_2(\hat{\mathcal{F}}_T, \mathcal{F}_u) \to 0$ almost everywhere, where d_2 is the Mallows metric defined on the space of distributions with finite variance.[7]

As discussed by Godfrey and Orme (1999), B artificial samples of size T can be generated from

$$y_t^* = x_t'\hat{\beta} + u_t^*, \quad t = 1, \ldots, T,$$

where $u_1^*, u_2^*, \ldots, u_T^*$ is a random sample drawn with replacement from $\hat{\mathcal{F}}_T$. If the artificially generated test statistics are denoted by $\tau_1^*, \tau_2^*, \ldots, \tau_B^*$, the p-value of $\tau_{(0)}$ can be estimated by

$$PV_{BS} = \frac{\sum_{b=1}^{B} \mathbf{1}\left(\tau_b^* > \tau_{(0)}\right)}{B}.$$

The null hypothesis is then rejected when $PV_{BS} \leq \alpha$, where α is the desired significance level.

3.2 MC tests

In contrast to the nonparametric bootstrap approach, Dufour *et al.* (2004, eq. 3) assume that the true error distribution is known. Let the cdf for ε used in the method of Dufour

[7] The metric d_2 is defined as follows. Let $\mathcal{F}_w$ and $\mathcal{F}_z$ be two distribution functions on the real line with $\int_{-\infty}^{\infty} |w|^2 d\mathcal{F}_w(w) < \infty$ and $\int_{-\infty}^{\infty} |z|^2 d\mathcal{F}_z(z) < \infty$, then

$$d_2(\mathcal{F}_w, \mathcal{F}_z) = \inf_{\mathcal{M}} \{E|w - z|^2\}^{1/2},$$

where $\mathcal{M}$ is the set of all joint distributions of w and z whose marginal distributions are $\mathcal{F}_w$ and $\mathcal{F}_z$, respectively; see Mallows (1972).

et al. (2004) be denoted by $\mathcal{G}$. In this case, for any given statistic τ whose distribution is independent of (β', σ^2), the MC approach generates test statistics which possess the same finite sample distribution, given the regressor values, as τ, provided $\mathcal{F} = \mathcal{G}$. That is, using iid drawings ε_t^+ with cdf $\mathcal{G}$, N samples can be generated as

$$y_t^+ = x_t'\hat{\beta} + \varepsilon_t^+, \quad t = 1, \ldots, T$$

which delivers test statistics, denoted $\tau_1^+, \tau_2^+, \ldots, \tau_N^+$. Notice that there is no need to scale ε_t^+ by $\hat{\sigma}$, due the invariance with respect to (β', σ^2). (Indeed, the scheme $y_t^+ = \varepsilon_t^+$ could be used, although this would create difficulties if, say, the test variables employed, z_t, contained squared predicted values.)

If $\mathcal{F} = \mathcal{G}$, $\tau_1^+, \ldots, \tau_N^+$ form a sample of iid random variables possessing the same finite sample distribution as τ. Thus $\left(\tau, \tau_1^+, \ldots, \tau_N^+\right)$ is a simple random sample of $N+1$ random variables, under the null, leading to the rejection rule: reject H_0 if $PV_{MC} \leq \alpha$, where

$$PV_{MC} = \frac{\sum_{i=1}^{N} \mathbf{1}\left(\tau_i^+ > \tau_{(0)}\right) + 1}{N+1}. \tag{13}$$

Under regularity conditions provided by Dufour *et al.* (2004), this rule provides an exact test when $\alpha\,(N+1)$ is an integer.

3.3 Robustness properties of simulation-based tests

In contrast to the exact procedures afforded by the MC approach when $\mathcal{F} = \mathcal{G}$, the non-parametric bootstrap approach only provides asymptotically valid inferences. However, it is important to examine the properties of the MC test procedure when $\mathcal{F} \neq \mathcal{G}$, i.e. under an incorrect choice of error distribution.[8] The examination of the asymptotic robustness of MC tests is based upon expansions of the type used by Beran (1988) to determine the orders of magnitude of ERP functions for asymptotic and bootstrap tests. It will be useful to start by outlining some of Beran's analysis using notation similar to that employed in his article.

[8] This issue is not pursued by Dufour *et al.* (2004), who assume that the correct choice is made.

Let the limit null cdf of τ under $\mathcal{F}$ be denoted by $H(z;\mathcal{F})$, with the finite sample null distribution function being $H_T(z;\mathcal{F}) = \Pr_T(\tau \leq z)$, where (in both) possible dependence on $\mathcal{F}$ is made explicit. It is assumed that the expansion given by Beran (1988) applies so that

$$H_T(z;\mathcal{F}) = H(z;\mathcal{F}) + T^{-j/2}h(z;\mathcal{F}) + O(T^{-(j+1)/2}), \tag{14}$$

uniformly in z, and for some integer $j \geq 1$ defined so that $h(z;\mathcal{F}) = O(1)$. Following Beran (1988), suppose that $H(\cdot;\mathcal{F})$ is continuous and strictly monotone over its support and that $h(\cdot;\mathcal{F})$ is continuous. For the bootstrap statistic τ^* generated under the law of $\hat{\mathcal{F}}_T$, the expansion corresponding to (14), which is for τ under $\mathcal{F}$, is

$$H_T\left(z;\hat{\mathcal{F}}_T\right) = H\left(z;\hat{\mathcal{F}}_T\right) + T^{-j/2}h\left(z;\hat{\mathcal{F}}_T\right) + O_p(T^{-(j+1)/2}). \tag{15}$$

When τ is not an asymptotic pivot, the limit null distribution depends upon $\mathcal{F}$ as described above. In this case the ERP of the asymptotic test, which can be carried out by rejecting if $H\left(\tau_{(0)};\hat{\mathcal{F}}_T\right) > 1-\alpha$, is $O(T^{-j_a/2})$, for some $j_a \leq j$; see Beran (1988, p. 691). By consideration of various expansions of the same general type, Beran (1988) shows that the ERP of the bootstrap test is also $O(T^{-j_a/2})$ in this case. If, on the other hand, τ is an asymptotic pivot, so that $H(z;\mathcal{F}) = H\left(z;\hat{\mathcal{F}}_T\right) = H(z)$, Beran (1998, p . 690) shows that the bootstrap test now has an ERP which is $O(T^{-(j+1)/2})$ whilst the asymptotic test has an ERP which is $O(T^{-j/2})$. (In the case of asymptotically pivotal chi-square statistics, $j = 2$.) Therefore, the nonparametric bootstrap delivers asymptotically valid inferences regardless of whether the test uses an asymptotic pivot, without any further assumptions about the form of $\mathcal{F}$, and is predicted to give better control over significance levels than first order asymptotic theory when the test statistic is an asymptotic pivot.

As noted above, if $\mathcal{F}$ is known, the MC methods of Subsection 3.2 can be used to obtain exact tests. Consider now, though, the MC test procedure under $\mathcal{G}$. Under this choice of error law, the finite sample null cdf of τ^+ can be expanded as

$$H_T(z;\mathcal{G}) = H(z;\mathcal{G}) + T^{-j/2}h(z;\mathcal{G}) + O(T^{-(j+1)/2}), \tag{16}$$

provided $\mathcal{G}$ satisfies the regularity conditions. If the test statistic is an asymptotic pivot, its limit null distribution is independent of the error law, so $H(z;\mathcal{F}) = H(z;\hat{\mathcal{F}}_T) = H(z;\mathcal{G}) = H(z)$. In particular, the simulated statistics $\tau_1^+, \tau_2^+, \ldots, \tau_N^+$ (generated under the artificial process associated with $\mathcal{G}$) and the actual statistic τ (under the true distribution associated with $\mathcal{F}$) have the same limit null cdf, namely $H(z)$. Since $\tau_1^+, \tau_2^+, \ldots, \tau_N^+$ and τ are asymptotically iid, the MC rejection rule (13) is asymptotically valid, with an ERP the same order in T as that for the asymptotic test; i.e., $O\left(T^{-j/2}\right)$.

If τ is not asymptotically pivotal, $H(z;\mathcal{F}) \neq H(z;\mathcal{G})$. Consequently $\tau_1^+, \tau_2^+, \ldots, \tau_N^+$ do not constitute an asymptotically valid reference set for τ because they do not possess the same limit distribution as τ. Therefore the MC rejection rule (13) can lead to asymptotically invalid inferences.

The above asymptotic analysis provides the following conclusions:

1. If τ is an asymptotic pivot:

   (a) the MC test procedure with correct choice of $\mathcal{G}$ ($\mathcal{G} = \mathcal{F}$) delivers exact inference in finite samples (the ERP is zero);

   (b) the MC test procedure with incorrect choice of $\mathcal{G}$ ($\mathcal{G} \neq \mathcal{F}$) has an ERP which is of the same order in T as the asymptotically valid test procedure which employs $H(z)$ as the reference distribution, and therefore delivers asymptotically valid inferences;

   (c) the nonparametric bootstrap test procedure also delivers asymptotically valid inferences, but has an ERP which is of smaller order in T than that of the asymptotically valid test procedure which employs $H(z)$ as the reference distribution.

2. If τ is not an asymptotic pivot:

   (a) the MC test procedure with correct choice of $\mathcal{G}$ ($\mathcal{G} = \mathcal{F}$) still delivers exact inference in finite samples (the ERP is zero);

(b) the MC test procedure with incorrect choice of $\mathcal{G}$ ($\mathcal{G} \neq \mathcal{F}$) has an ERP which is $O(1)$ and therefore delivers asymptotically invalid inferences;

(c) the nonparametric bootstrap test procedure still delivers asymptotically valid inferences, but has an ERP which is of the same order in T as that of the asymptotically valid test procedure which employs $H(z; \hat{\mathcal{F}}_T)$ as the reference distribution.

Thus, for example, the MC-based Koenker test with any choice of $\mathcal{G}$ (satisfying the appropriate regularity conditions discussed in Section 2) is asymptotically valid, but asymptotic theory predicts that the nonparametric bootstrap version has smaller ERP in finite samples. On the other hand, the MC version of the Breusch-Pagan test with incorrect choice of $\mathcal{G}$ is invalid both asymptotically and in finite samples, whilst the MC version of the Glesjer test is asymptotically valid when $\mathcal{G} \neq \mathcal{F}$, provided both $\mathcal{G}$ and $\mathcal{F}$ are symmetric.

Therefore, the value of the MC approach is that it has the potential to provide valid exact inferences for all test procedures, but only if the correct choice of $\mathcal{G}$ is made. Since, in general, the true error distribution is unknown, this is a very strong assumption. If $\mathcal{G} \neq \mathcal{F}$, then it still provides asymptotically valid inferences but only for tests based on pivotal statistics. In contrast, the nonparametric bootstrap, whilst not providing exact inferences, always yields asymptotically valid inferences for all unknown $\mathcal{F}$ satisfying the regularity conditions.

In the light of the above discussion, it is important to compare and contrast the finite sample performances of MC and nonparametric bootstrap test procedures in order to provide applied workers with general guidance concerning the efficacy of the two approaches under incomplete information concerning the error distribution. This is provided in Section 5, which reports the results of a Monte Carlo study described in the next Section.

4 Monte Carlo design

To facilitate comparison with the results reported by Dufour *et al.* (2004), finite sample significance levels are investigated using their Monte Carlo data generation process which

can be written as

$$y_t = \sum_{j=1}^{6} x_{tj}\beta_j + u_t, \;\; u_t \;\; iid(0, \sigma^2), \quad t = 1, \ldots, T, \tag{17}$$

in which: $x_{t1} = 1$, so that β_1 is an intercept term; the regressor values $x_{t2}, \ldots, x_{t6}$ are independent drawings from the uniform distribution $U(0, 10)$; $\beta_j = 1$ for all j; and $T = 50, 100$. In practical situations, the regressors may not be approximately uniformly distributed and, since the design matrix can have an impact on the finite sample performance of tests, it seems useful to conduct a second set of experiments.

The second set of experiments uses the same basic regression model as the first set, i.e. (17). However, the regressors are taken from a data set provided by Greene (2003, Table F6.1), rather than being obtained from pseudo-random number generators. This data set contains 27 cross-section observations on the following production-activity variables: VA, a measure of value added; LAB, a measure of labour input; and CAP, an index of capital stock. These variables are used to construct the $k = 6$ regressors for (17) with $x_{t1} = 1$, $x_{t2} = \log(LAB_t)$, $x_{t3} = \log(CAP_t)$, $x_{t4} = x_{t2}^2$, $x_{t5} = x_{t3}^2$, and $x_{t6} = x_{t2}x_{t3}$ for $t = 1, \ldots, T$. The values of the parameters of (17) in experiments that use Greene's data are the corresponding OLS estimates that he reports for a regression in which $\log(VA)$ is the dependent variable; see Greene (2003, p. 103). In order to obtain samples sizes similar to those in the first set of experiments, the genuine observations for $T = 27$ are reused to obtain $T = 54$ and $T = 108$, according to

$$x_{tj} = x_{t+27,j} = x_{t+54,j} = x_{t+81,j} \quad \text{for } t = 1, \ldots, 27 \text{ and } j = 1, \ldots, 6.$$

For both sets of experiments, the specification of regressor values and model coefficients allows the calculation of conditional mean values $E(y_t|x_t)$, $t = 1, \ldots, T$, that are fixed over replications. The addition of a pseudo-random error to the mean function gives an artificial observation: all random number generators are taken from the NAG library and are used in FORTRAN programs. The choice of distributions to be employed for drawing error

terms is based upon established usage. More specifically, the following distributions are often used in Monte Carlo studies and satisfy the regularity requirements given by Koenker (1981): Normal; Student $t(5)$; Uniform; $\chi^2(2)$; and Lognormal. In addition to these five distributions, errors are also drawn using a pseudo-random number generator for the Cauchy distribution. This last distribution does not have finite moments and so does not satisfy the usual regularity conditions. It is included because it is the subject of investigation and comment by Dufour *et al.* (2004).

Having combined errors and means to obtain an artificial sample of T observations, (17) can be estimated by OLS and tests of the assumption of homoskedasticity can be carried out. The following eight test statistics are examined.

(a) *Goldfeld-Quandt test with natural ordering of data*

Using the notation of Section 2 above, this test is implemented in the first set of experiments by setting $T_1 = 2T/5$, $T_2 = T/5$, and $T_3 = 2T/5$ for $T = 50, 100$. In the second set of experiments, the corresponding values are, as suggested by Johnston and DiNardo (1997, p. 168), $T_j = T/3$ for $j = 1, 2, 3$ and $T = 54, 108$. The test statistic obtained with the natural ordering of the data is denoted by GQ_n.

(b) *Goldfeld-Quandt test with ordering of data by values of squared OLS predictions*

For this test, data are reordered according to the rank numbers of the squared predicted values $\hat{y}_t^2$. After this reordering, the Goldfeld-Quandt test is applied using the same combinations of T_1, T_2 and T_3 as are used for test (a). The statistic calculated after ordering by the ranks of the terms $\hat{y}_t^2$ is denoted by GQ_o.

It should be noted that the Goldfeld-Quandt tests GQ_n and GQ_o are not included because it is believed that there will often be sufficient information to permit an ordering of variances under heteroskedasticity. Instead, these tests are used to provide some evidence about the effects of nonnormality on GQ_o, relative to GQ_n, when it cannot be assumed that OLS predicted and residual values are independent; see the discussion in Subsection 2.2.5.

(c) *Breusch-Pagan test*

The specific version of the general test statistic derived by Breusch and Pagan (1979), under the assumption of normality, is computed from the artificial regression of the scaled squared OLS residuals $\hat{u}_t^2/\hat{\sigma}^2$ on the regressors of (17). This statistic is denoted by BP_x.

(d) *Glejser test*

The statistic for the Glejser test is, following Dufour *et al.* (2004), the conventional F-statistic for testing that all slope coefficients equal zero in the artificial regression of $|\hat{u}_t|$ on the regressors of (17). The asymptotic properties of this test are discussed in Section 2 above. The test statistic is denoted by G_x.

(e) *Koenker test (first version)*

Koenker's (1981) procedure is used in two forms. The first version has as its test statistic T times the R^2 from the OLS regression of $\hat{u}_t^2$ on the regressors of (17). This test statistic is denoted by K_x.

(f) *Koenker test (second version)*

The second version of a Koenker-type check is a modification of White's (1980) general test for heteroskedasticity. The test statistic is T times the R^2 from the OLS regression of $\hat{u}_t^2$ on the regressors and nonredundant squared regressors of (17). (The modification of White's strategy is, therefore, that cross-products of regressors are not considered.) This statistic is denoted by K_W.

(g) *MSSI modified Glejser test (first version)*

The modified version of Glejser's test that is proposed by MSSI is, like Koenker's test, implemented using two versions; see Section 2 for details of the modification and its purpose. In the first version, the test variables are the regressors of (17) and the associated statistic is denoted by $MSSI_x$.

(h) *MSSI modified Glejser test (second version)*

The second version of the *MSSI* procedure uses the same test variables as K_W and leads to a test statistic denoted by $MSSI_W$.

The nominal significance level for all of the tests is, as in the experiments of Dufour *et al.* (2004), set equal to 5%. Estimated rejection frequencies are derived from 25000 replications. For each test statistic (a)-(h), several evaluations of statistical significance are made: theory-based critical values, the MC approach of Dufour *et al.* (2004), and the nonparametric bootstrap recommended by Godfrey and Orme (1999) are all employed. Theory-based critical values are not always even asymptotically valid in the experiments but are as follows: the GQ_n and GQ_o statistics are assessed by reference to the $F(T_3 - 6, T_1 - 6)$ distributions; G_x is compared with critical values from $F(5, T - 6)$ distributions; BP_x, K_x, and $MSSI_x$ all use the $\chi^2(5)$ distribution; and, in the first set of experiments, the common reference distribution for K_W and $MSSI_W$ is $\chi^2(10)$, while for the second group of experiments, in which Greene's data are used, this distribution is $\chi^2(8)$.

When MC test techniques are employed, $N = 99$ replications are used with all of the six possible error distributions. Thus, for any given correct choice of the error distribution, there are also five incorrect models being used. The six sets of estimates produced provide evidence on the robustness of the MC method advocated by Dufour *et al.* (2004). As an alternative to using a parametric setting, the nonparametric bootstrap is implemented, as in Godfrey and Orme (1999), with $B = 400$ bootstraps.

5 Monte Carlo results

Before looking at the results from the Monte Carlo study, it is important to define criteria to evaluate the performance of the different tests considered. Given the large number of replications performed, the standard asymptotic test for proportions can be used to test hypotheses about the true significance levels. Since some of the tests studied in these experiments are exact, we can expect that the null hypothesis that their rejection frequencies equal the nominal significance level of 5% is accepted in most cases. In these experiments, this null hypothesis is accepted (at the 5% level) for estimated rejection frequencies in the range 4.73% to 5.27%. In practice, however, what is important is not that the significance

level of the test is identical to the chosen nominal size, but rather that the true and nominal rejection frequencies stay reasonably close, even when the test is only approximately valid. Cochran (1952) suggested that a test can be regarded as robust relative to a nominal value of 5% if its actual significance level is between 4% and 6%. Considering the number of replications used in these experiments, estimated rejection frequencies within the range 3.75% to 6.30% are viewed as providing evidence consistent with the robustness of the test, according to Cochran's definition.

Tables 1 to 4 display a set of selected results of the simulation experiments. To economize on space, only a representative sample of results is provided. For each of the two regressor sets, estimates are reported for three error distributions: normal, a symmetric nonnormal distribution, and an asymmetric distribution. With the design of Dufour *et al.* (2004), the error laws are normal, $t(5)$ and $\chi^2(2)$, and, when Greene's (2003) data are used, the distributions are normal, uniform, and lognormal. In these Tables, the results of the MC tests obtained using the correct distributions are presented in bold face and the results obtained using nonparametric bootstrap critical values are in italic.

Estimates for Cauchy errors are not reported in Tables 1-4. While the Cauchy distribution has importance in various areas of econometrics, it is not clear that it is a useful choice for the error model in the context of studying tests of the assumption of homoskedasticity in regression models. If the errors were to be Cauchy, the mean and variance functions of y_t in (17), conditional upon any set of finite values of the x_{tj}, would not exist. Therefore, although included in the simulation experiments, the case of Cauchy errors is not used to illustrate general findings. However, the evidence that is obtained by using Cauchy errors can be easily summarized: not surprisingly, given the failure to satisfy regularity conditions, only the MC tests using the correct choice of Cauchy errors perform well.[9]

[9] The results for Cauchy errors, along with those for all other error distributions, can be obtained from the authors upon request.

Table 1: Rejection frequencies at the 5% level (Dufour *et al.* design, $n = 50$)

Errors:	Critical values	GQ_n	GQ_o	BP_x	G_x	K_x	K_W	$MSSI_x$	$MSSI_W$
Normal	χ^2 or F	4.96	4.98	4.54	5.18	4.06	3.51	3.82	3.08
	NP Bootstrap	*4.96*	*5.17*	*4.12*	*5.72*	*5.23*	*5.38*	*5.27*	*5.39*
	MC/normal	**4.78**	**4.76**	**4.90**	**5.02**	**5.12**	**5.08**	**5.12**	**5.06**
	MC/$t(5)$	1.84	1.75	10.61	5.36	5.99	5.55	5.28	5.43
	MC/uniform	9.51	9.60	1.46	5.05	4.55	4.91	4.95	5.06
	MC/$\chi^2(2)$	0.69	1.38	22.94	13.45	6.92	6.82	5.62	6.06
	MC/lognormal	0.01	0.12	49.03	20.32	8.87	7.33	6.52	6.80
$t(5)$	χ^2 or F	10.49	10.21	22.01	5.37	3.88	3.79	3.82	3.33
	NP Bootstrap	*6.52*	*6.22*	*2.86*	*5.40*	*4.67*	*5.06*	*4.97*	*5.06*
	MC/normal	9.74	9.96	2.12	4.31	4.11	4.42	4.55	4.56
	MC/$t(5)$	**4.78**	**4.86**	**4.82**	**4.98**	**4.81**	**4.88**	**4.98**	**4.84**
	MC/uniform	15.62	15.33	0.65	4.47	3.95	4.24	4.64	4.77
	MC/$\chi^2(2)$	2.48	3.89	11.35	12.29	5.70	5.93	5.15	5.54
	MC/lognormal	0.19	0.66	28.41	19.51	7.68	6.52	6.37	6.16
$\chi^2(2)$	χ^2 or F	15.12	20.19	41.98	21.31	6.62	5.32	7.28	5.10
	NP Bootstrap	*7.33*	*6.42*	*2.26*	*4.55*	*4.36*	*4.76*	*4.98*	*5.00*
	MC/normal	14.94	10.74	0.91	1.66	3.48	3.59	4.38	4.25
	MC/$t(5)$	8.86	6.07	2.02	1.84	4.02	3.94	4.70	4.26
	MC/uniform	20.66	15.63	0.27	1.64	3.23	3.33	4.36	4.14
	MC/$\chi^2(2)$	**5.30**	**5.01**	**4.86**	**4.93**	**4.92**	**4.79**	**4.83**	**4.93**
	MC/lognormal	0.70	1.02	14.04	8.06	6.25	5.49	5.76	5.52

Table 2: Rejection frequencies at the 5% level (Dufour *et al.* design, $n = 100$)

Errors:	Critical values	GQ_n	GQ_o	BP_x	G_x	K_x	K_W	$MSSI_x$	$MSSI_W$
Normal	χ^2 or F	5.08	5.17	4.81	5.18	4.37	4.11	4.47	4.01
	NP Bootstrap	*5.60*	*5.19*	*4.53*	*5.47*	*5.23*	*5.32*	*5.22*	*5.16*
	MC/normal	**5.38**	**4.87**	**4.93**	**4.95**	**5.02**	**5.12**	**4.98**	**4.98**
	MC/$t(5)$	1.20	0.94	13.83	5.06	5.76	5.72	4.96	5.14
	MC/uniform	12.16	12.00	1.07	5.04	4.71	4.71	4.94	4.85
	MC/$\chi^2(2)$	0.32	0.49	29.95	13.96	6.48	6.62	5.30	5.49
	MC/lognormal	0.00	0.00	68.48	21.65	8.02	7.70	5.94	6.26
$t(5)$	χ^2 or F	12.69	12.18	31.66	5.37	3.93	4.15	4.40	4.40
	NP Bootstrap	*6.10*	*6.34*	*3.22*	*5.12*	*4.91*	*4.96*	*4.88*	*5.09*
	MC/normal	12.35	12.54	1.66	4.40	4.35	4.20	4.52	4.72
	MC/$t(5)$	**5.14**	**5.28**	**4.85**	**4.73**	**5.01**	**4.94**	**4.82**	**5.00**
	MC/uniform	20.53	20.83	0.34	4.35	4.17	3.89	4.42	4.72
	MC/$\chi^2(2)$	2.46	3.46	11.96	13.24	5.66	5.59	4.80	5.26
	MC/lognormal	0.12	0.28	36.60	20.45	6.94	6.85	5.66	6.00
$\chi^2(2)$	χ^2 or F	17.36	20.99	53.78	24.57	5.62	4.92	6.50	5.25
	NP Bootstrap	*6.59*	*6.40*	*2.76*	*4.81*	*4.60*	*4.96*	*4.98*	*5.04*
	MC/normal	17.36	14.51	0.58	1.47	3.72	3.81	4.69	4.46
	MC/$t(5)$	8.78	7.04	1.86	1.53	4.26	4.18	4.66	4.49
	MC/uniform	25.04	21.42	0.13	1.54	3.58	3.56	4.45	4.30
	MC/$\chi^2(2)$	**5.13**	**5.10**	**4.80**	**4.90**	**4.82**	**5.11**	**4.85**	**5.02**
	MC/lognormal	0.32	0.43	19.12	8.00	5.95	6.01	5.62	5.55

Table 3: Rejection frequencies at the 5% level (Greene's design, $n = 54$)

Errors:	Critical values	GQ_n	GQ_o	BP_x	G_x	K_x	K_W	$MSSI_x$	$MSSI_W$
Normal	χ^2 or F	4.84	4.81	4.47	5.66	4.68	5.13	4.27	4.31
	NP Bootstrap	*5.07*	*5.37*	*4.25*	*5.63*	*5.25*	*5.34*	*5.16*	*5.22*
	MC/normal	**4.93**	**5.08**	**4.83**	**5.10**	**4.92**	**4.89**	**4.96**	**5.06**
	MC/$t(5)$	1.82	1.91	10.87	5.00	4.26	3.27	4.74	4.40
	MC/uniform	9.92	9.36	1.59	5.93	6.33	8.22	5.87	6.56
	MC/$\chi^2(2)$	0.67	8.08	23.66	11.62	4.20	2.93	4.61	4.48
	MC/lognormal	0.01	4.84	51.38	14.19	3.61	1.81	3.87	3.26
Uniform	χ^2 or F	1.72	2.04	0.06	8.20	4.66	3.92	5.91	4.64
	NP Bootstrap	*4.23*	*4.24*	*6.02*	*5.14*	*4.77*	*4.61*	*4.58*	*4.93*
	MC/normal	1.93	2.11	13.42	3.94	3.72	2.92	4.04	3.87
	MC/$t(5)$	0.53	0.57	26.24	3.90	3.15	1.74	3.80	3.36
	MC/uniform	**5.12**	**4.90**	**4.83**	**4.77**	**4.99**	**4.97**	**4.76**	**5.12**
	MC/$\chi^2(2)$	0.11	4.27	47.82	9.94	3.13	1.76	3.62	3.38
	MC/lognormal	0.01	2.09	79.25	11.70	2.62	0.97	3.05	2.46
Lognormal	χ^2 or F	25.00	50.05	65.11	28.74	12.64	14.14	11.42	11.73
	NP Bootstrap	*7.92*	*5.95*	*1.39*	*4.36*	*6.05*	*6.36*	*5.16*	*5.72*
	MC/normal	23.70	5.24	0.17	1.13	7.97	12.25	6.65	7.73
	MC/$t(5)$	17.79	3.31	0.45	1.14	6.29	8.40	6.24	6.65
	MC/uniform	29.00	7.45	0.05	1.57	10.86	17.67	8.03	9.64
	MC/$\chi^2(2)$	12.98	7.59	1.39	3.81	6.36	7.88	5.90	7.05
	MC/lognormal	**4.80**	**6.14**	**4.91**	**4.83**	**4.94**	**4.90**	**4.82**	**4.91**

Table 4: Rejection frequencies at the 5% level (Greene's design, $n = 108$)

Errors:	Critical values	GQ_n	GQ_o	BP_x	G_x	K_x	K_W	$MSSI_x$	$MSSI_W$
Normal	χ^2 or F	4.85	4.84	4.84	5.37	4.62	4.73	4.71	4.44
	NP Bootstrap	*5.26*	*5.31*	*4.47*	*5.54*	*5.17*	*5.41*	*5.31*	*5.32*
	MC/normal	**4.95**	**5.05**	**5.00**	**5.17**	**4.84**	**5.09**	**5.11**	**4.98**
	MC/$t(5)$	1.13	1.16	14.23	5.20	4.34	3.48	5.00	4.83
	MC/uniform	12.01	11.69	1.10	5.46	5.20	5.95	5.38	5.39
	MC/$\chi^2(2)$	0.28	3.76	30.67	13.46	4.46	3.69	5.00	5.19
	MC/lognormal	0.00	1.02	70.14	17.67	3.97	2.12	4.55	4.11
Uniform	χ^2 or F	1.12	1.09	0.02	6.75	4.96	4.64	5.88	5.26
	NP Bootstrap	*4.40*	*4.63*	*6.30*	*5.39*	*5.37*	*5.25*	*5.12*	*5.26*
	MC/normal	1.22	1.34	19.41	4.80	4.78	4.22	4.84	4.74
	MC/$t(5)$	0.10	0.12	41.95	4.89	4.34	3.04	4.77	4.53
	MC/uniform	**5.13**	**5.12**	**5.18**	**5.10**	**5.22**	**5.15**	**5.06**	**5.07**
	MC/$\chi^2(2)$	0.01	0.75	68.24	12.36	4.49	3.15	4.79	4.76
	MC/lognormal	0.00	0.18	95.92	16.22	3.84	1.72	4.28	3.75
Lognormal	χ^2 or F	28.72	50.25	82.48	32.57	11.28	12.81	9.78	10.04
	NP Bootstrap	*7.10*	*7.46*	*2.01*	*4.91*	*5.90*	*6.18*	*5.34*	*5.51*
	MC/normal	28.94	10.35	0.11	1.06	6.52	10.92	5.53	6.02
	MC/$t(5)$	21.08	6.46	0.31	1.10	5.78	7.88	5.46	6.00
	MC/uniform	34.94	13.83	0.02	1.24	7.06	12.66	5.86	6.64
	MC/$\chi^2(2)$	16.18	10.01	0.96	3.53	6.07	8.27	5.37	6.20
	MC/lognormal	**5.06**	**6.52**	**5.08**	**4.90**	**4.77**	**4.90**	**4.89**	**4.88**

The results in Tables 1-4 show that, when critical values are taken from either χ^2 or F distributions, asymptotically valid tests are not always reliable in finite samples, and that, not surprisingly, estimates for asymptotically invalid tests are not close to 5%. Both of these features are illustrated in Table 1 by results for errors derived from the $\chi^2(2)$ distribution: the estimate for the asymptotically valid $MSSI_x$ test is 7.28%; and the estimate for the asymptotically invalid BP_x test is 41.98%.

Given the evidence of the inadequacy of asymptotic critical values, which corroborates that reported by Godfrey and Orme (1999), the simulation-based methods of Section 3 are of interest. Consider first the cases in which MC tests are carried out with the correct error distribution. Except for combinations with GQ_o tests and nonnormal errors, the results of Dufour *et al.* (2004) imply exact validity of MC tests for such cases. As expected, with 25000 replications, the corresponding estimates are usually observed to be close to 5%. The estimates for GQ_o tests with the correct choice of nonnormal error distribution are also quite close to 5%, except when Greene's data provide regressor values and the errors are lognormal. Overall, with the right error model, the MC tests, as anticipated from the arguments of Dufour *et al.* (2004), do well. However, the main purpose of our experiments is to gauge the robustness of the MC tests to departures from the hypothesized error distribution. Therefore, it is interesting to focus on the results for MC tests when the distribution used to compute the critical values (or p-values) is different from the true distribution of the errors. When discussing the estimates relevant to the issue of robustness, it is again useful to distinguish between those test statistics that are asymptotically pivotal and those that are not.

For the asymptotic pivots, viz. K_x, K_W, $MSSI_x$ and $MSSI_W$, the MC method yields asymptotically valid, but not exact, tests when the wrong error model is selected; see Subsection 3.3. The nonparametric bootstrap versions of test are not only asymptotically valid, but also enjoy a refinement in the ERP. The results obtained in our experiments, as illustrated by the relevant parts of Tables 1-4, indicate that, despite being asymptotically valid, MC tests based on the wrong error law often have estimates that fail to comply with

Cochran's criterion of robustness, as set out above. On the other hand, the nonparametric bootstrap forms of K_x, K_W, $MSSI_x$ and $MSSI_W$ are well-behaved and it is noteworthy that they are not markedly inferior to MC tests that use the correct error distribution.

The statistics that are not asymptotically pivotal are BP_x, G_x, GQ_n and GQ_o. The results show how combining these statistics with an incorrect error law in a MC test can produce large discrepancies between estimated rejection rates and 5%. Since, as argued in Section 3, such combinations imply a lack of asymptotic validity, these discrepancies are not unexpected. It is particularly interesting to consider the behaviour of the MC version of the Glejser test under misspecification of the error distribution. Dufour *et al.* (2004) comment that the "estimation effect" problem that afflicts the Glejser test (see Godfrey, 1996) is irrelevant when their MC method is used. While this comment is certainly true when the error distribution is correctly specified, the results reported here clearly show that, as a result of the estimation effects problem, the size of the Glejser test is distorted when the degree of skewness of the distribution is misspecified. Specifically, the test systematically underrejects (overrejects) the null when the assumed errors are less (more) skewed that the true errors. (Recall, however, that the results in Section 3.3 predict that the MC version of the Glesjer test will be asymptotically valid, despite misspecification of the error distribution, provided both the true and assumed distributions are symmetric, and satisfy the regularity conditions.) In contrast, the kurtosis of the distribution has no significant effect on the behaviour of the test. Therefore, the estimation effect noted by Godfrey (1996) is critical for the performance of the MC version of the Glejser test, under an incorrect choice of the error distribution, and determines the direction of the bias of the estimated significance level.[10]

The nonparametric bootstrap variants of (at least) BP_x, G_x and GQ_n are asymptotically valid, but the evidence suggests that these procedures are not reliable in finite samples of the magnitudes considered; see, for example, the results for BP_x in Table 1.

[10] Although it does not suffer from the estimation effects problem, similar results are found for the BP test due to the incorrect estimation of the variance of u_t^2 under non-normality. In this case it is the difference between the kurtosis of the true and assumed distributions that is important for the performance of the test.

Overall there are two recommendations to applied workers that emerge from the Monte Carlo study. First, avoid the use of test statistics that are not asymptotically pivotal: use Koenker's Studentized score, rather than the Breusch-Pagan version, and use the MSSI modification of Glejser's test, rather the original procedure. If there is information about the alternative that suggests a GQ-type test would be feasible, use the test variable identified in this alternative in a Koenker-type check. The Koenker test involving the OLS regression of $\hat{u}_t^2$ on an intercept term and $\hat{y}_t^2$ provides an example of such an alternative to GQ; see Godfrey and Orme (1999) for results on the finite sample behaviour of the former test. Second, having selected an asymptotically pivotal test statistic, use the nonparametric bootstrap to assess the statistical significance of its sample value.

6 Conclusions

The main advantage of the MC procedure recently proposed by Dufour *et al.* (2004) to obtain tests for heteroskedasticity is that, under parametric distributional assumptions, it leads to simple exact tests, even when the test statistic used has an unknown asymptotic distribution. This approach not only gives perfect control over the empirical significance level of the tests, but also increases the number of different tests available to practitioners. Based on simulation experiments, Dufour *et al.* (2004) find that, indeed, their MC procedure leads to tests with very good behaviour under the null and that the new tests it makes available have good power properties. However, the authors do not evaluate the sensitivity of their tests to departures from incidental assumptions on the shape of the error distribution. This paper reassesses the usefulness of the MC tests for heteroskedasticity, focusing on the robustness of this procedure to departures from the maintained distributional assumptions.

Dufour *et al.* (2004) show that, under correct distributional assumptions, all standard heteroskedasticity tests are pivotal in finite samples. However, this result critically depends on the validity of the hypothesized error distribution since the distribution of the test statistics depends on its shape, at least in finite samples. When the assumed distribution of the

error is incorrect, the MC-based tests are no longer exact, and their asymptotic validity depends on whether or not the tests are based on asymptotically pivotal statistics. Whenever they are based upon statistics that are not asymptotic pivots, the asymptotic validity of the MC tests recommended by Dufour *et al.* (2004) cannot be guaranteed under misspecification of the assumed distribution.

The results in Section 3 imply that MC tests based on asymptotically pivotal statistics have large sample validity, even if the error distribution is misspecified. However, the results of our simulation experiments suggest that, even with a correctly specified error distribution, the MC tests do not have an important advantage over analogous tests based on critical values obtained by nonparametric bootstrap, at least for sample sizes of practical interest. Moreover, under incorrect distributional assumptions, MC tests based on asymptotic pivots are often outperformed by the corresponding tests based on the nonparametric bootstrap, where the latter have an error in rejection probability that is of smaller order in T than those of the MC and asymptotic tests.

Overall, although the results of Dufour *et al.* (2004) are certainly interesting and potentially useful in cases where the researcher has confidence in the maintained distributional assumptions, the use of the MC tests for heteroskedasticity they suggest cannot generally be recommended for the more standard situation in which little is known about the distribution governing the errors of the model.[11] It is instead recommended that the nonparametric bootstrap be used when assessing the statistical significance of checks for heteroskedasticity. If the applied researcher is tackling a more difficult problem of, e.g., looking at the minimum of a set of p-values, as in the new tests given by Dufour *et al.* (2004), a double bootstrap can be used to control finite sample significance levels. The use of double bootstrap schemes with non-standard tests for heteroskedasticity is an interesting topic for future research; see

[11] Our arguments do not, of course, apply to the MC test of normality proposed by Dufour, Farhat, Gardiol and Khalaf (1998).

Godfrey (2003) for an application to the minimum p-value of several tests of the regression mean function.

References

Beran, R., 1988. Prepivoting test statistics: a bootstrap view of asymptotic refinements. Journal of the American Statistical Association 83, 687-697.

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

Cochran, W.G., 1952. The χ^2 test of goodness of fit. Annals of Mathematical Statistics 23, 315-345.

Cook, R.D., Weisberg, S., 1983. Diagnostics for heteroscedasticity in regression. Biometrika 70, 1-10.

Dufour, J.-M., Farhat, A., Gardiol, L., Khalaf, L., 1998. Simulation-based finite sample normality tests in linear regressions. The Econometrics Journal 1, 154-173.

Dufour, J.-M., Khalaf, L., Brenard, J.-T., Genest, I., 2004. Simulation-based finite-sample tests for heteroskedasticity and ARCH effects. Journal of Econometrics, forthcoming.

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

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

Godfrey, L.G., 1979. Testing for multiplicative heteroskedasticity. 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. Journal of Econometrics 72, 275-299.

Godfrey, L.G., 2003. Controlling the overall significance level of a battery of least squares diagnostic tests. Unpublished paper, University of York.

Godfrey, L.G., Orme, C.D., 1999. The robustness, reliability, and power of heteroskedasticity tests. Econometric Reviews 18, 169-194.

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

Greene, W., 2003. Econometric analysis, 5th ed. Prentice-Hall, Englewood Cliffs, NJ.

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

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

Horowitz, J.L. 1994. Bootstrap-based critical values for the information matrix test, Journal of Econometrics 61, 395-411.

Horowitz, J.L., Savin, N.E., 2000. Empirically relevant critical values for hypothesis tests: a bootstrap approach. Journal of Econometrics 95, 375-389.

Im, K.S., 2000. Robustifying the Glejser test of heteroskedasticity. Journal of Econometrics 97, 179-188.

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

Machado, J.A.F., Santos Silva, J.M.C., 2000. Glejser's test revisited, Journal of Econometrics 97, 189-202.

Mallows, C.L., 1972. A note on asymptotic joint normality. Annals of Mathematical Statistics 43, 508-515.

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.

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

White, H., 1980. A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. Econometrica 48, 817-838.