Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Statistics & Data Analysis 49 (2005) 361 – 376

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap

Emmanuel Flachaire*

EUREQua, Université Paris 1 Panthéon-Sorbonne, Maison des Sciences Economiques, 106-112 bd de l'Hopital, Paris Cedex 1375647, France

Available online 19 June 2004

Abstract

In regression models, appropriate bootstrap methods for inference robust to heteroskedasticity of unknown form are the wild bootstrap and the pairs bootstrap. The finite sample performance of a heteroskedastic-robust test is investigated with Monte Carlo experiments. The simulation results suggest that one specific version of the wild bootstrap outperforms the other versions of the wild bootstrap and of the pairs bootstrap. It is the only one for which the bootstrap test always gives better results than the asymptotic test.

Keywords: Wild bootstrap; Pairs bootstrap; Heteroskedasticity-robust test; Monte Carlo simulations

1. Introduction

Let us consider the linear heteroskedastic model

$$y_t = X_t\beta + u_t, \quad E(u_t|X_t) = 0, \quad E(u_t^2|X_t) = \sigma_t^2, \tag{1}$$

where y_t is a dependent variable, X_t an exogenous k-vector, β and σ_t^2 are the unknown parameters and unknown variances of the error term, respectively, and u is error term. Inference on the parameters requires special precautions when the error terms u_t are heteroskedastic. Then, the OLS estimator of the covariances of the estimates of $\hat{\beta}$ are in general biased and inconsistent, and so conventional tests are not t and F distributions, even asymptotically.

* Tel.: +33-1-4407-8214; fax: +33-1-4407-8231

E-mail address: emmanuel.flachaire@univ-paris1.fr (E. Flachaire).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.05.018

This problem is solved in Eicker (1963) and White (1980), where a *Heteroskedasticity Consistent Covariance Matrix Estimator*, or HCCME, is proposed that permits asymptotically correct inference on β in the presence of heteroskedasticity of unknown form

$$(X^\top X)^{-1} X^\top \hat{\Omega} X (X^\top X)^{-1}, \tag{2}$$

where the $n \times n$ diagonal matrix $\hat{\Omega}$ has elements $a_t^2 \hat{u}_t^2$, where $\hat{u}_t$ is the OLS residual. MacKinnon and White (1985) considers a number of possible forms of HCCME and shows that, in finite samples, they too, can be seriously biased, as also t or F statistics based on them, especially in the presence of observations with high leverage; see also Chesher and Jewitt (1987), who show that the extent of the bias is related to the structure of the regressors. We refer to the basic version of the HCCME, proposed in Eicker (1963) and White (1980) as HC_0 ($a_t = 1$) and to the other forms considered by MacKinnon and White (1985) as HC_1, HC_2 and HC_3. Specifically,

$$HC_1 : a_t = \sqrt{\frac{n}{n-k}}, \quad HC_2 : a_t = \frac{1}{\sqrt{1-h_t}}, \quad HC_3 : a_t = \frac{1}{1-h_t}, \tag{3}$$

where $h_t \equiv X_t (X^\top X)^{-1} X_t^\top$ is the tth element of the orthogonal projection matrix on to the span of the columns of X. MacKinnon and White (1985) and Chesher and Jewitt (1987) show that, in terms of error in the rejection probability (ERP), HC_0 is outperformed by HC_1, which is in turn outperformed by HC_2 and HC_3. The last two cannot generally be ranked, although HC_3 has been shown in a number of Monte Carlo experiments to be superior in typical cases. However, even if HC_2 and HC_3 perform better in finite samples, ERP is still significant if the sample size is not large. Then, it makes sense to consider whether bootstrap methods might be used to alleviate their size distortion.

Bootstrap methods normally rely on simulation to approximate the finite-sample distribution of test statistics under the null hypotheses they test. For such methods to be reasonably accurate, it is desirable that the data-generating process (DGP) used for drawing bootstrap samples be as close as possible to the DGP that generated the observed data, assuming that the DGP satisfies the null hypothesis. This presents a problem if the null hypothesis admits heteroskedasticity of unknown form: If the form is unknown, it cannot be imitated in the bootstrap DGP.

A technique used to overcome this difficulty is the so-called wild bootstrap, developed in Liu (1988) following suggestions in Wu (1986) and Beran (1986). Liu establishes the ability of the wild bootstrap to provide refinements for the linear regression model with heteroskedastic errors. Further evidence is provided in Mammen (1993). Both Liu and Mammen show, under a variety of regularity conditions, that the wild bootstrap is asymptotically justified, in the sense that the asymptotic distribution of various statistics is the same as the asymptotic distribution of their wild bootstrap counterparts. They also show that, in some circumstances, asymptotic refinements are available that lead to agreement higher than leading order between the asymptotic distributions of the raw and bootstrap statistics. Recently, Davidson and Flachaire (2001) proposes a procedure with better finite-sample performance than the version usually recommended in the literature and give exact results for some specific cases. Experimental results in Godfrey and Orme (2001) indicate that tests should be implemented using this procedure.

Another very popular technique that has been used to overcome the problem of heteroskedasticity of unknown form is the so-called pairs bootstrap, or bootstrap by pairs, proposed in Freedman (1981). In its original form, the pairs bootstrap is implemented by randomly resampling the data directly with replacement. Monte Carlo evidence suggests that inference based on this procedure is not always accurate (Horowitz, 2000). However, an improved version of the pairs bootstrap is proposed in Mammen (1993), and subsequently modified in Flachaire (1999), in which a resampling scheme is defined that respects the null hypothesis and simulation results show that its performance is highly improved but the size distortion is still significant.

The important question of whether the error in the rejection probability of a test based on the HCCME is smaller with the wild bootstrap than with the pairs bootstrap has been studied in a few experiments, in MacKinnon (2002) Brownstone and Valletta (2001) and Horowitz (2000). Here, we provide additional evidence for this question and for power comparisons through Monte Carlo experiments with different versions of the pairs bootstrap and of the wild bootstrap.

In Section 2, I present the wild bootstrap and the pairs bootstraps. The model design is described in Section 3, and simulation results are presented in Section 4. Section 5 concludes.

2. Bootstrap methods

Numerical results show that hypothesis tests based on asymptotic HCCME can be seriously misleading with small samples, see for instance MacKinnon and White (1985), Davidson and Flachaire (2001) and Godfrey and Orme (2001). It then makes sense to consider bootstrap methods to improve reliability of such tests.

In regression models, the principle of the bootstrap can be expressed as follows:

To compute a test, the bootstrap principle is to construct a data-generating process, called the bootstrap DGP, based on estimates of the unknown parameters and probability distribution. The distribution of the test statistic under this artificial DGP is called the bootstrap distribution. A *P*-value bootstrap can be calculated using the bootstrap distribution as the nominal distribution.

It is often impossible to calculate the bootstrap distribution analytically: we approximate it through simulations (bootstrap resampling). It is clear that, if the bootstrap DGP is close to the true DGP, the bootstrap distribution should be close to the true distribution.

Theoretical developments show that under some regularity conditions, the bootstrap error of the rejection probability, or bootstrap ERP, converges more quickly to zero than the asymptotic ERP. For fixed sample size n, the convergence rate of the ERP of a test statistic based on its asymptotic distribution is in general of order n^{-1}. Beran (1988) shows that bootstrap inference is refined with order $n^{-1/2}$ when the quantity bootstrapped is asymptotically pivotal and when the parameters and distribution estimates in the bootstrap DGP are consistent. Davidson and MacKinnon (1999) shows that a further refinement, in general of order $n^{-1/2}$, can be obtained if the bootstrapped test statistic and the DGP bootstrap are asymptotically independent. These two successive refinements, of order n^{-1}, give a bootstrap ERP faster convergence than an asymptotic ERP.

Let us consider the linear regression model

$$y_t = x_{1t}\beta + X_{2t}\gamma + u_t \tag{4}$$

in which the explanatory variables are assumed to be strictly exogenous, in the sense that, for all t, x_{1t} and X_{2t} are independent of all of the error terms u_s, $s=1,\dots,n$: $E(u_t|x_1, X_2)=0$. The row vector X_{2t} contains observations on $k-1$ variables, of which, if $k>1$, one is a constant. We test the null hypothesis that the coefficient β of the first regressor x_{1t} is equal to β_0, with a t test statistic

$$\tau = \frac{\hat\beta - \beta_0}{\sqrt{\widehat{\mathrm{Var}}(\hat\beta)}}, \tag{5}$$

where $\hat\beta$ is the OLS parameter estimate of β and $\widehat{\mathrm{Var}}(\hat\beta)$ is the heteroskedasticity consistent variance estimate of $\hat\beta$. We can compute an asymptotic test, using the asymptotic χ^2 distribution to compute a P value, or a bootstrap test, using a bootstrap distribution to compute a P value. Many different bootstrap tests can be computed with different bootstrap methods.

2.1. *Pairs bootstrap*

The bootstrap by pairs, proposed in Freedman (1981), consists in resampling the regressand and regressors together from the original data: a bootstrap sample $(y^\star, x_1^\star, X_2^\star)$ is obtained by drawing raws independently with replacement from the matrix (y, x_1, X_2). It is clear that the bootstrap sample does not come from a parametric model respecting the null hypothesis. Thus, we need to modify the test statistic to be compatible with the data, see Hall (1992). The distribution of the modified test statistic,

$$\tau = \frac{\hat\beta^\star - \hat\beta}{\sqrt{\widehat{\mathrm{Var}}(\hat\beta^\star)}} \tag{6}$$

is the pairs bootstrap distribution, where $\hat\beta^\star$ is the OLS estimate from a bootstrap sample.

Flachaire (1999) proposes modifying the resampling scheme so that the null hypothesis is respected in the bootstrap DGP. The principle is to consider the following bootstrap DGP:

$$y_t^\star = x_{1t}^\star \beta_0 + X_{2t}^\star \tilde\gamma + u_t^\star, \tag{7}$$

where $\tilde\gamma$ is the OLS restricted parameter estimate, and $(x_{1t}^\star, X_{2t}^\star, u_t^\star)$ is a $k+1$ vector drawn from the matrix $(x_1, X_2, a\hat u)$, where $a\hat u$ is a vector with typical element $a_s\hat u_s$, $\hat u_s$ is an OLS unrestricted residual and a_s can take the forms HC_0, HC_1, HC_2 or HC_3. Then, a bootstrap sample $(y^\star, x_1^\star, X_2^\star)$ is calculated based on the bootstrap DGP (7). It is clear that this last bootstrap DGP respects the null. Then, the distribution of the test statistic

$$\tau = \frac{\hat\beta^\star - \beta_0}{\sqrt{\widehat{\mathrm{Var}}(\hat\beta^\star)}} \tag{8}$$

is the pairs bootstrap distribution, where $\hat\beta^\star$ is the OLS estimate from a bootstrap sample.

2.2. Wild bootstrap

The wild bootstrap is developed in Liu (1988) following suggestions in Wu (1986) and Beran (1986). To generate a bootstrap sample, we use the following bootstrap DGP:

$$y_t^\star = x_{1t}\,\beta_0 + X_{2t}\tilde{\gamma} + a_t \tilde{u}_t\,\varepsilon_t^\star, \tag{9}$$

where $\tilde{u}_t$ is the OLS restricted residual and $\varepsilon_t^\star$ is white noise following a distribution, F, with expectation $E(\varepsilon_t^\star) = 0$ and variance $E(\varepsilon_t^{\star 2}) = 1$. The distribution of the test statistic

$$\tau = \frac{\hat{\beta}^\star - \beta_0}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}^\star)}} \tag{10}$$

is the wild bootstrap distribution, where $\hat{\beta}^\star$ is the OLS estimate from a bootstrap sample. In the literature, the further condition that $E(\varepsilon_t^{\star 3}) = 1$ is often added. Liu (1988) considers model (4) with $k = 1$ and shows that, with the extra condition, the first three moments of the bootstrap distribution of an HCCME-based statistic are in accord with those of the true distribution of the statistic up to order n^{-1}. Mammen (1993) suggests what is probably the most popular choice for the distribution of the $\varepsilon_t^\star$, namely the following two-point distribution:

$$F_1: \quad \varepsilon_t^\star = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } p = (\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2 & \text{with probability } 1-p. \end{cases} \tag{11}$$

Recently, Davidson and Flachaire (2001) have shown that the Rademacher distribution

$$F_2: \quad \varepsilon_t^\star = \begin{cases} 1 & \text{with probability } 0.5, \\ -1 & \text{with probability } 0.5 \end{cases} \tag{12}$$

always gives better results than the version usually recommended in the literature F_1, and gives exact results for some specific cases. They show that the rate of convergence of the error in the rejection probabilities, or ERP, is at most $n^{-3/2}$ with symmetric errors and $n^{-1/2}$ with asymmetric errors. Rather than limiting the analysis to a comparison of the order of the leading term in the expansion, they consider the Edgeworth expansion in its entirety through order n^{-1} in order to understand finite sample behavior better. They show that the full Edgeworth expansion of the wild bootstrap with F_2 is smaller than that with F_1 if the sample size is small enough, and also when high leverage observations are present. Their simulation results along with those of Godfrey and Orme (2001) indicate that this version is better than others and should be preferred in practice.

2.3. Wild bootstrap vs. Pairs bootstrap

In the light of the bootstrap principle, we can compare the wild bootstrap and the pairs bootstrap in regression models.

Since regressors are drawn in the pairs bootstrap resampling scheme, the regressors in the pairs bootstrap DGP are not exogenous and since we draw regressors and residuals in the same time, once $X_t^\star$ is known so is $u_t^\star$, so $E(u_t^\star|X_t^\star) \neq 0$ (for instance, $E(u_t^\star|X_t^\star = X_i) = \sum_{t=1}^n \hat{u}_t \; P(\hat{u}_t|X_t^\star = X_i) = \hat{u}_i$). In the true DGP, if regressors X are assumed to be exogenous and $E(u_t|X_t) = 0$, then we can see that these two basic hypothesis are not respected in the pairs bootstrap DGP. The bootstrap principle suggests getting a bootstrap DGP as close as possible to the true DGP. In our case, we could improve the performance of the pairs bootstrap with the following modifications:

(1) $E(u_t^\star|X_t^\star) = 0$ is respected if we draw $(X^\star, u^\star)$ in $(X, \hat{u})$ and if we consider the following bootstrap DGP: $y_t^\star = X_t^\star \hat{\beta} + u_t^\star \varepsilon_t^\star$, where $\varepsilon_t^\star$ are mutually independent drawings from some auxiliary distribution such that $E(\varepsilon_t^\star) = 0$ and $E(\varepsilon_t^{\star 2}) = 1$.

(2) $X^\star$ is exogenous if we do not resample regressors. Note that we cannot resample residuals independently of regressors because heteroskedasticity could be a function of regressors. This leads us to consider the following bootstrap DGP: $y_t^\star = X_t \hat{\beta} + \hat{u}_t \varepsilon_t^\star$. This last DGP is the *wild* bootstrap DGP.

Thus, by imposing the two hypotheses to the pairs bootstrap DGP to get a DGP closer to the true DGP, we are led to consider the wild bootstrap DGP. We would anticipate that bootstrap tests based on the wild bootstrap would give a better numerical performance than those based on the pairs bootstrap. This is borne out by simulation results in the following sections.

3. Model design

In the experiments, I consider the linear regression model

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \sigma_t \varepsilon_t, \tag{13}$$

where both regressors are drawn from the standard lognormal distribution and the true parameters are $\beta_0 = \beta_1 = \beta_2 = 0$. For heteroskedastic data $\sigma_t^2 = x_{t1}^2$, and ε_t is white noise following $N(0, 1)$. The sample size is $n = 100$, the number of Monte-Carlo simulations is equal to $N = 10,000$, and the number of bootstrap replications is equal to $B = 499$. We test the null hypothesis $H_0 : \beta_1 = 0$ with a t statistic based on the HCCME,

$$\tilde{\tau} = x_1^\top M_2 y / (x_1^\top M_2 \tilde{\Omega} M_2 x_1)^{1/2}, \tag{14}$$

where $\tilde{\Omega}$ is an $n \times n$ diagonal matrix with diagonal elements $(a_t \tilde{u}_t)^2$, a transformation of the tth *restricted* residual $\tilde{u}_t$ from the restricted regression

$$y_t = \beta_0 + \beta_2 x_{2t} + \sigma_t \varepsilon_t \tag{15}$$

and $M_2 = I - X_2(X_2^\top X_2)^{-1} X_2^\top$ is the orthogonal projection matrix on to the orthogonal complement of the span of the columns of $X_2 = [\imath \; x_2]$ where $\imath$ is a vector unity. An estimator used more frequently in practice replace $\tilde{\Omega}$ with $\hat{\Omega}$, having diagonal elements $(a_t \hat{u}_t)^2$, a transformation of the tth *unrestricted* residual $\hat{u}_t$ from regression (13). Note that simulation results do not depend on the choices of β_0, β_2, and on the scales of regressors and σ_t^2, because (14) does not depend on those parameters. The model design is chosen to make heteroskedasticity robust tests difficult: heteroskedasticity is made a function of the regressors and, because of the lognormal distribution, a few extreme observations are often present among the regressors.

The wild bootstrap DGP is

$$y_t^\star = \tilde{u}_t (1 - h_t)^{-1/2} \, \varepsilon_t^\star. \tag{16}$$

Note that, since the distribution of the τ statistic we consider is independent of the parameters β_0 and β_2 of the regression ($M_2 x_2 = M_2 \iota = 0$), we may set $\beta_0 = \beta_2 = 0$ in the bootstrap DGP without loss of generality. In our simulations, we consider four different cases: F_1 or F_2 is the distribution of $\varepsilon_t^\star$ in the bootstrap DGP and the statistic is computed with the restricted $\tilde{\tau}$ or with the unrestricted residuals $\hat{\tau}$.

The pairs bootstrap DGP proposed in Freedman (1981) consists in resampling directly in (y, x_1, x_2). In this case, the bootstrap statistic tests the modified null hypothesis $\beta_1 = \hat{\beta}_1$. The version proposed in Flachaire (1999) consists in generating the bootstrap sample $(y^\star, x_1^\star, x_2^\star)$ by resampling $(x_{1t}^\star, x_{2t}^\star, u_t^\star)$ in the matrix with elements $(x_{1t}, x_{2t}, a_t \hat{u}_t)$, where $a_t = (1 - h_t)^{-1/2}$ and $\hat{u}_t$ is the OLS unrestricted residual, and generating $y_t^\star$ with the DGP

$$y_t^\star = \tilde{\beta}_0 + \tilde{\beta}_2 x_{2t}^\star + u_t^\star. \tag{17}$$

In that case, the bootstrap statistic tests the null hypothesis $\beta_1 = 0$. In the simulations, we consider four cases: the pairs bootstrap that respects the null hypothesis H_0 or the pairs bootstrap that respects the alternative H_1 and statistics computed with the restricted $\tilde{\tau}$ or with the unrestricted residuals $\hat{\tau}$.

4. Experimental results

In this section, I study the finite-sample behavior of asymptotic and bootstrap tests robust to heteroskedasticity of unknown form. To be useful, a test must be able to discriminate between the null hypothesis and the alternative. In the one hand, a test is reliable if it rejects the null at the nominal level α, when the null hypothesis is true. Otherwise, significant ERP should be exhibited. On the other hand, a test is powerful if it rejects strongly the null, when the null hypothesis it is not true.

- ERP results of the experiments are shown using the graphical *P* value discrepancy plots, as described in Davidson and MacKinnon (1998). These figures show,as a function of the nominal level α, the difference between the true rejection probability and the nominal level, that is to say the error in the rejection probability, or ERP. This is also called size distortion.
- Power results of the experiments are shown using Power functions. It is not desirable to plot power against nominal level to compare the power of alternative test statistics if all the tests exhibit significant ERP. Then, power functions plot power against true level, as defined in the Size-Power curves proposed in Davidson and MacKinnon (1998). This is often called "level-adjusted" power. Note that the simulation results are not sensitive to the choice of β_0 and β_2 in the DGP under the alternative, because statistic (14) does not depend on those parameters. Thus, DGPs under the alternative are defined with different values of β_1 and it is not necessary to define a *drifting* DGP as in Davidson and MacKinnon (2002).

I investigate finite-sample performance of asymptotic and bootstrap HCCME-based test statistics, computed both with the restricted and unrestricted residuals, for the following cases:

$asymp$	: asymptotic test.
$wboot1$	: wild bootstrap with F_1, Mammen (1993).
$wboot2$	: wild bootstrap with F_2, Davidson and Flachaire (2001).
$pboot1$	: pairs bootstrap under H_1, Freedman (1981).
$pboot0$	: pairs bootstrap under H_0, Flachaire (1999).

4.1. Which transformation of residuals: HC_0, HC_1, HC_2 or HC_3?

MacKinnon and White (1985), Chesher and Jewitt (1987) and Long and Ervin (2000) show that the error of the rejection probability of an asymptotic test based on HC_0 is larger than HC_1, which is in turn larger than those of HC_2 and HC_3. The last two cannot be ranked in general, although HC_3 has been shown in a number of Monte Carlo experiments to be superior in typical cases. The Davidson and Flachaire (2001) experimental results show that the HC_3 version of the HCCME is better than the other three, and the best transformation a_t in the definition of the wild bootstrap DGP should be the same as that used for the HCCME. Similarly, the Flachaire (1999) experiments results show that the transformation HC_3 should be preferred to HC_0 with the pairs bootstrap. All this leads us for our simulations to use the HC_3 transformation in the bootstrap DGPs and to compute the HC_3 version of the HCCME.

4.2. Restricted or unrestricted residuals?

There has been some discussion in the literature on the use of restricted residuals rather than unrestricted residuals. Restricted residuals denotes the OLS estimates subject to the restriction that is being tested. On the one hand, if the null hypothesis is true, we can argue that the use of restricted residuals should improve the reliability of the tests. The Davidson and MacKinnon(1985) results suggest that *asymptotic* tests based on restricted residuals are more reliable—they slightly under-reject the null—while asymptotic tests based on unrestricted residuals largely over-reject the null hypothesis when true. Thus, they recommend the use of restricted residuals to compute the HCCME based test. On the other hand, van Giersbergen and Kiviet (2002) argue that if the null is not true, the use of unrestricted residuals could improve the power and should be preferred. However, this last argument is not supported in simulation experiments, see MacKinnon (2002). For the *bootstrap* test, simulation results in Davidson and Flachaire (2001) show that it is not very important whether one uses restricted or unrestricted residuals, but that it is a mistake to mix unrestricted residuals in the HCCME and restricted residuals for the bootstrap DGP. However, additional experiments in Godfrey and Orme (2001) show that it is not possible to have a good control of ERP by the use of the wild bootstrap if HCCME is computed with unrestricted residuals. There are a few results in favor of the use of restricted residuals, but

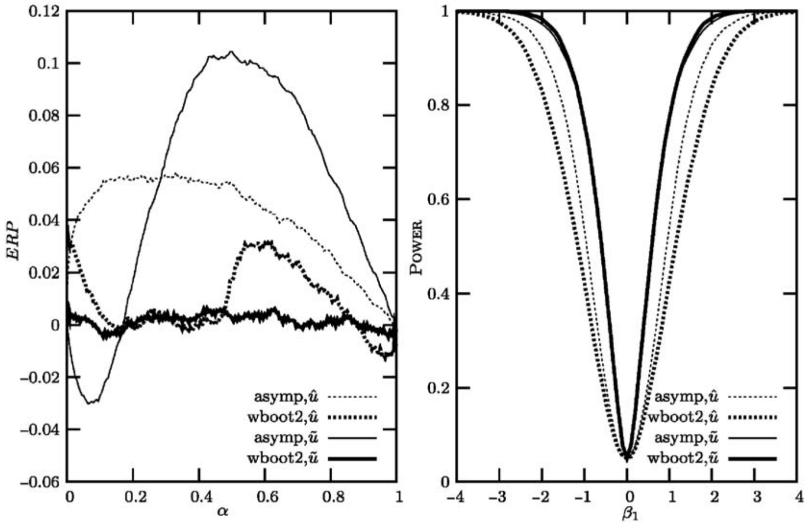

Fig. 1. ERP and POWER of *t*-test based on restricted or unrestricted residuals.

they are not very strong. We conduct some experiments to study this problem in our model design.

Fig. 1, on the left, shows error in the rejection probability, or ERP, of asymptotic $(asymp)$ and bootstrap $(wboot2)$ tests based on the HCCME computed with both *restricted* and *unrestricted* residuals as a function of the nominal level α. This P value discrepancy plot shows significant ERP for all tests, except for the bootstrap test based on restricted residuals and on the wild bootstrap with F_2 $(wboot2, \tilde{u})$. In practice, we are concerned with a small nominal level of $\alpha = 0.01$ or $\alpha = 0.05$. If we restrict our attention to small nominal levels, we can see that asymptotic tests based on restricted residuals $(asymp, \tilde{u})$ slightly under-reject the null (*ERP* slightly negative) and asymptotic tests based on unrestricted residuals $(asymp, \hat{u})$ over-reject the null. These results are similar to those of Davidson and MacKinnon(1985). It is interesting to examine all nominal levels: even if at small nominal levels, ERPs of $(asymp, \tilde{u})$ are smaller than those of $(asymp, \hat{u})$, it is not true for larger nominal levels. In other words, the asymptotic distributions of these two tests are not good approximations to the true distributions of the test statistic. Even if the use of restricted residuals gives slightly better results in some cases, it is not always true. We can also see from this figure, results similar to Godfrey and Orme (2001) about bootstrap tests: we do not have a good control over ERP when we use unrestricted residuals $(wboot2, \hat{u})$ and we have a very good control over it when we use restricted residuals $(wboot2, \tilde{u})$.

Fig. 1, on the right, shows the power of the asymptotic $(asymp)$ and bootstrap $(wboot2)$ tests based on HCCME computed with both *restricted* and *unrestricted* residuals, at a rejection probability level $RP = 0.05$, against different DGPs under the alternative hypothesis β_1. Under the alternative, a DGP is defined with $\beta_0 = \beta_2 = 0$ and $\beta_1 = -4, -3.9, \ldots, 3.9, 4$. The power increases as β_1 goes away from 0, and if $\beta_1 = 0$ the power is equal to the rejection probability level 0.05. Then, the most powerful test would reject the null when $\beta_1 \neq 0$ and would take the form ⊤ in our figure. We see that the asymptotic and bootstrap tests based on restricted residuals have similar power, but this is not true for those based on unrestricted residuals. This result is anticipated by the theoretical results in Davidson and MacKinnon (2002), where it is shown that the difference between the powers of the asymptotic and bootstrap tests should be of the same magnitude as the difference between their ERPs. We also see an additional interesting result: tests based on restricted residuals exhibit more power than those based on unrestricted residuals.

Finally, these results suggest that the use of restricted residuals is highly to be recommended in practice: ERPs of HCCME based tests can be controlled by the use of the wild bootstrap and the tests are more powerful. In the following experiments, we make use of restricted residuals.

4.3. Wild bootstrap vs. Pairs bootstrap: base case

MacKinnon (2002) investigates the performance of the pairs and wild bootstraps using a bootstrap DGP with restricted residuals and a HCCME-based *t* test with unrestricted residuals. We have seen above the importance of using HCCME-based *t* test with restricted rather than unrestricted residuals. In particular, some of our experiments, along with Godfrey and Orme (2001), show that the wild bootstrap does not perform similarly if we use unrestricted residuals. Consequently, it is of interest to investigate the performance of the wild and pairs bootstrap for HCCME based tests computed with restricted residuals, when comparing these two distinct bootstrap methods.

Fig. 2, on the left, shows the ERP of asymptotic and bootstrap HCCME based tests computed with restricted residuals. From this figure it is clear that only the wild bootstrap F_2 version of the *t* test $(wboot2)$ performs well and avoids significant ERP at all nominal levels. If we restrict our attention to the nominal level $\alpha = 0.05$, the pairs bootstrap proposed by Freedman (1981) $(pboot1)$ appears to perform well: its ERP is not far away from 0. However, if we consider all nominal levels, its behavior is not good and is quite similar to that of the asymptotic test. Once again, we see the importance of considering more than one nominal level: we can conclude that $wboot2$ performs well, not $pboot1$ and the other tests. Moreover, except for the test based on the wild bootstrap proposed in Davidson and Flachaire (2001), it is not possible to conclude from this figure that the other bootstrap schemes $(wboot1$, $pboot1$ and $pboot2)$ give better results than the asymptotic test $(asymp)$.

Fig. 2, on the right, shows the Power of asymptotic and bootstrap HCCME based tests computed with restricted residuals at a rejection probability level $RP = 0.05$. We see that the wild bootstrap tests $(wboot1$ and $wboot2)$ and the asymptotic tests $(asymp)$ have similar power. This is an additional interesting result: the pairs bootstrap computed under the null $(pboot2)$ displays a slight loss of power and the pairs bootstrap computed under the alternative $(pboot1)$, as proposed by Freedman (1981), displays a large loss of power.

4.4. High leverage observations

Chesher and Jewitt (1987) shows that the HCCME finite sample performance is related to high leverage observations in the data, that is, unbalanced regressors. Our model design

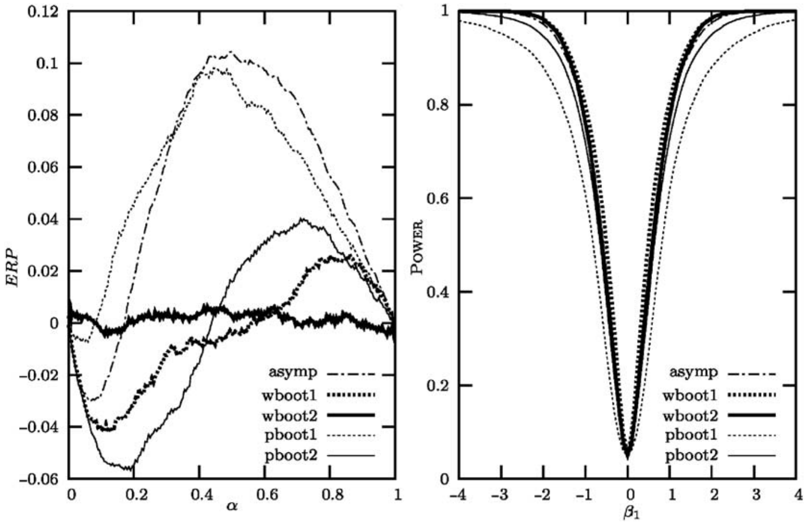

Fig. 2. ERP and POWER of *t*-test, wild vs. pairs bootstrap: base case.

Table 1
ERPs of *t*-test with different sample size *n*, $\alpha = 0.05$

n	asymp	wboot1	wboot2	pboot1	pboot2
50	−0.024	−0.024	0.002	0.001	−0.030
100	−0.027	−0.031	0.002	−0.007	−0.032
200	−0.022	−0.029	−0.001	−0.011	−0.032
300	−0.048	−0.040	−0.005	−0.048	−0.050
400	−0.046	−0.038	−0.002	−0.043	−0.050
500	−0.047	−0.038	−0.001	−0.042	−0.049
1000	−0.045	−0.036	−0.001	−0.032	−0.045

is chosen to include high leverage observations: regressors are drawn from the lognormal distribution so a few observations are often quite extreme. To reduce the level of high leverage observations, I conduct a first experiment increasing the sample size and a second experiment with regressors drawn from the Normal distribution. Table 1 shows the ERP of the asymptotic and bootstrap tests, as in the base case, with different sample size, at the nominal level $\alpha = 0.05$. We see that the convergence to zero of the ERP of the asymptotic test is very slow and only the ERP of the bootstrap test $wboot2$ is always close to zero.

Fig. 3 shows the ERP of the asymptotic and bootstrap tests, as in the base case, except that the regressors are drawn from the standard Normal distribution $N(0, 1)$ and the sample size is reduced to $n = 50$. The conclusions are similar to the base case (Fig. 2, left); that is, $wboot2$ performs very well. However, all ERPs are largely reduced. In addition, the ERP of the asymptotic test with Normal regressors and $n = 100, 200, 300, 400, 500$ are,

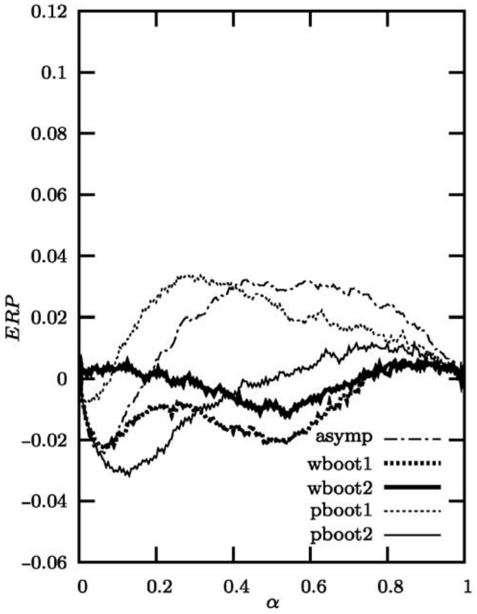

Fig. 3. No high leverage, $n = 50$.

respectively

$$\mathrm{ERP} = -0.010,\ \ -0.013,\ \ -0.004,\ \ -0.001,\ \ -0.001, \tag{18}$$

at nominal level $\alpha = 0.05$. It is interesting to see that the HCCME-based tests perform better with a small sample and well-balanced regressors when compared to a large sample and unbalanced regressors. These results highlight the fact that the finite sample performance of the HCCME-based tests is more sensitive to the structure of the regressors than to the sample size.

4.5. Asymmetric error term

Davidson and Flachaire (2001) shows that the rate of convergence of the ERP is at most $n^{-3/2}$ if we use the wild bootstrap with F_2 ($wboot2$) and if the error terms are *symmetric*. In comparison, the rate of convergence of the ERP is at most n^{-1} if we use the wild bootstrap with F_1 ($wboot1$) and if the error terms are *symmetric*. Then, under the assumption of symmetry, we can expect that $wboot2$ will outperform $wboot1$. In addition, Davidson and Flachaire (2001) shows that, under the assumption of *asymmetric* error terms, the ERP is at most of order $n^{-1/2}$ with $wboot2$ and of order n^{-1} with $wboot1$. However, based on formal Edgeworth expansions, they show that $wboot1$ does not outperform $wboot2$ if the sample size is small and if there are high leverage observations: it corresponds to cases where heteroskedasticity gives serious problems. Their simulation results, along with

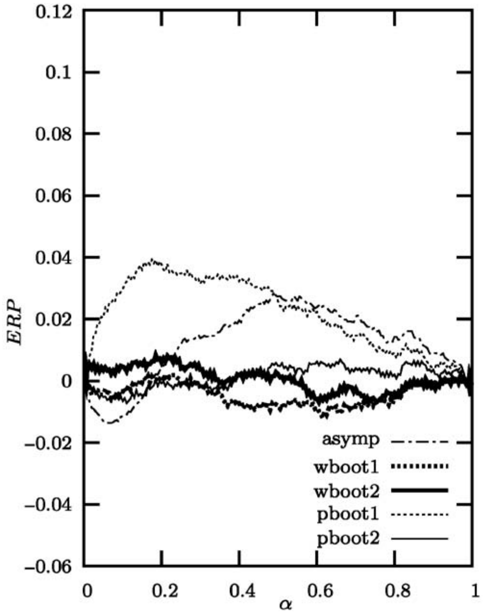

Fig. 4. Asymmetric errors, $\chi^2(5)$.

those in Godfrey and Orme (2001), suggest that even if errors are not symmetric, $wboot2$ should be preferred in practice. I check this last result with one additional experiment. Fig. 4 shows the ERP of the asymptotic and bootstrap tests, as in the base case, except that errors are drawn from a Chi-square distribution $\chi^2(5)$. From this figure, we see that $wboot1$ and $wboot2$ perform well.

4.6. F-test

Here we are interested in a null hypothesis with more than one restriction. We test the null that all coefficients except the slope are equal to zero, $\beta_1 = \beta_2 = 0$, with an F-test. Fig. 5 shows the ERP of the asymptotic and bootstrap tests, as in the base case, except that we use a F-test statistic. Once more, only the test based on the wild bootstrap with F_2 $(wboot2)$ performs well. The other tests exhibit significant ERPs with larger magnitudes (see the y-axis scale) than if we had used a t-test (Fig. 2, left).

4.7. Homoskedastic error term

What is the penalty for using the wild and the pairs bootstraps when the errors are homoskedastic and inference based on assuming the conventional t statistic is reliable? To answer this question, Fig. 6 shows the ERP of the asymptotic and bootstrap tests, as in the base case, except that error terms are homoskedastic $\sigma_t^2 = 1$. We can see that the

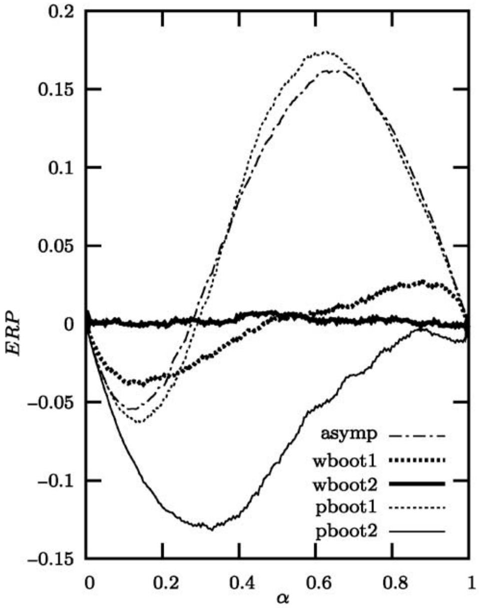

Fig. 5. *F*-test.

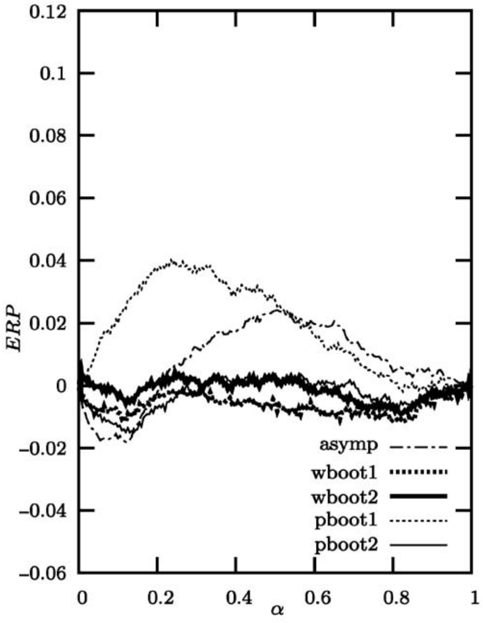

Fig. 6. Homoskedastic errors.

ERPs are largely reduced compared to the heteroskedastic base case (Fig. 2, left). Once again, $wboot2$ performs better than the others and the penalty attached to using this wild bootstrap version is very small.

5. Conclusion

I examine finite-sample performances of heteroskedasticity-robust tests. Simulation results initially suggest computing heteroskedasticity-robust tests with restricted residuals rather than unrestricted residuals to achieve a gain in power. Additional experiments show, however, that the version of the wild bootstrap proposed in Davidson and Flachaire (2001) always gives better numerical performance than the pairs bootstrap and another version of the wild bootstrap proposed in the literature: ERP is always very small and Power is similar to that of the asymptotic test. Further, the results show that this version of the wild bootstrap is the only one for which the bootstrap test outperforms the asymptotic test in all the cases, a property that is not true for other versions of the wild and pairs bootstraps.

References

Beran, R., 1986. Discussion of “Jackknife bootstrap and other resampling methods in regression analysis” by C.F.J. Wu. Ann. Statist. 14, 1295–1298.

Beran, R., 1988. Prepivoting test statistics: a bootstrap view of asymptotic refinements. J. Amer. Statist. Assoc. 83 (403), 687–697.

Brownstone, D., Valletta, R., 2001. The bootstrap and multiple imputations: harnessing increased computing power for improved statistical tests. J. Econom. Perspectives 15, 129–142.

Chesher, A., Jewitt, I., 1987. The bias of a heteroskedasticity consistent covariance matrix estimator. Econometrica 55, 1217–1222.

Davidson, R., Flachaire, E., 2001. The wild bootstrap, tamed at last. Working paper, IER#1000, Queen’s University.

Davidson, R., MacKinnon, J.G., 1985. Heteroskedasticity-robust tests in regression directions. Ann. de l’INSEE 59/60, 183–218.

Davidson, R., MacKinnon, J.G., 1998. Graphical methods for investigating the size and power of hypothesis tests. The Manchester School 66 (1), 1–26.

Davidson, R., MacKinnon, J.G., 1999. The size distortion of bootstrap tests. Econometric Theory 15, 361–376.

Davidson, R., MacKinnon, J.G., 2002. The power of bootstrap and asymptotic tests. Unpublished paper, revised November.

Eicker, B., 1963. Limit theorems for regression with unequal and dependant errors. Ann. Math. Statist. 34, 447–456.

Flachaire, E., 1999. A better way to bootstrap pairs. Econom. Lett. 64, 257–262.

Freedman, D.A., 1981. Bootstrapping regression models. Ann. Statist. 9, 1218–1228.

Godfrey, L.G., Orme, C.D., 2001. Significance levels of heteroskedasticity-robust tests for specification and misspecification: some results on the use of wild bootstraps. Paper presented at ESEM’2001, Lausanne.

Hall, P., 1992. The Bootstrap and Edgeworth Expansion. Springer Series in Statistics. Springer, New York.

Horowitz, J.L., 2000. The bootstrap. (Eds.), in: .in: J.J.,Heckman,E.E.,Leamer,(Eds.), Handbook of Econometrics, Vol. 5. Elsevier Science, Amsterdam.

Liu, R.Y., 1988. Bootstrap procedure under some non-i.i.d. models. Ann. Statist. 16, 1696–1708.

Long, J.S., Ervin, L.H., 2000. Heteroscedasticity consistent standard errors in the linear regression model. Amer. Statist. 54, 217–224.

MacKinnon, J.G., 2002. Bootstrap inference in econometrics. Canad. J. Econom. 35, 615–645.

MacKinnon, J.G., White, H.L., 1985. Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. J. Econometrics 21, 53–70.

Mammen, E., 1993. Bootstrap and wild bootstrap for high dimensional linear models. Ann. Statist. 21, 255–285.

van Giersbergen, N.P.A., Kiviet, J.F., 2002. How to implement bootstrap hypothesis testing in static and dynamic regression model: test statistic versus confidence interval approach. J. Econometrics 108, 133–156.

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

Wu, C.F.J., 1986. Jackknife bootstrap and other resampling methods in regression analysis. Ann. Statist. 14, 1261–1295.