The Wild Bootstrap, Tamed at Last

by

Russell Davidson

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13236 Marseille cedex 02, France

Department of Economics
McGill University
Montreal, Quebec, Canada
H3A 2T7

email: **Russell.Davidson@mcgill.ca**

and

Emmanuel Flachaire

Université Paris I Panthéon-Sorbonne
Maison des Sciences Économiques
106-112 bd de l'Hôpital
75647 Paris Cedex 13

email: **emmanuel.flachaire@univ-paris1.fr**

Abstract

Various versions of the wild bootstrap are studied as applied to regression models with heteroskedastic disturbances. We show that, in one very specific case, perfect bootstrap inference is possible, and a substantial reduction in the error in the rejection probability of a bootstrap test is available much more generally. However, the version of the wild bootstrap with this desirable property does not benefit from the skewness correction afforded by the most popular version of the wild bootstrap in the literature. Simulation experiments are used to show why this defect does not prevent the preferred version from having the smallest error in rejection probability in small and medium-sized samples. It is concluded that this version should be used in practice.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are very grateful to James MacKinnon for helpful comments on an earlier draft, to participants at the ESRC Econometrics Conference (Bristol), especially Whitney Newey, and to several anonymous referees. Remaining errors are ours.

April 2007

1. Introduction

Inference on the parameters of the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u},$$

where $\boldsymbol{y}$ is an n-vector containing the values of the dependent variable, $\boldsymbol{X}$ an $n \times k$ matrix of which each column is an explanatory variable, and $\boldsymbol{\beta}$ a k-vector of parameters, requires special precautions when the disturbances $\boldsymbol{u}$ are heteroskedastic, a problem that arises frequently in work on cross-section data. With heteroskedastic disturbancess, the usual OLS estimator of the covariance of the OLS estimates $\hat{\boldsymbol{\beta}}$ is in general asymptotically biased, and so conventional t and F tests do not have their namesake distributions, even asymptotically, under the null hypotheses that they test. The problem was solved by Eicker (1963) and White (1980), who proposed a heteroskedasticity consistent covariance matrix estimator, or HCCME, that permits asymptotically correct inference on $\boldsymbol{\beta}$ in the presence of heteroskedasticity of unknown form.

MacKinnon and White (1985) considered a number of possible forms of HCCME, and showed that, in finite samples, they too, as also t or F statistics based on them, can be seriously biased, 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. But since, unlike conventional t and F tests, HCCME-based tests are at least asymptotically correct, it makes sense to consider whether bootstrap methods might be used to alleviate their small-sample size distortion.

Bootstrap methods normally rely on simulation to approximate the finite-sample distribution of test statistics under the null hypotheses they test. In order for such methods to be reasonably accurate, it is desirable that the data-generating process (DGP) used for drawing bootstrap samples should be as close as possible to the true DGP that generated the observed data, assuming that that 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.

In the face of this difficulty, the so-called wild bootstrap was developed by Liu (1988) following a suggestion of Wu (1986) and Beran (1986). Liu established the ability of the wild bootstrap to provide refinements for the linear regression model with heteroskedastic disturbances, and further evidence was provided by Mammen (1993), who showed, under a variety of regularity conditions, that the wild bootstrap, like the (y, X) bootstrap proposed by Freedman (1981), 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. These authors also show that, in some circumstances, asymptotic refinements are available, which lead to agreement between the distributions of the raw and bootstrap statistics to higher than leading order asymptotically.

In this paper, we consider a number of implementations both of the Eicker-White HCCME and of the wild bootstrap applied to them. We are able to obtain one

exact result, where we show that, in the admittedly unusual case in which the hypothesis under test is that all the regression parameters are zero (or some other given fixed vector of values), one version of the wild bootstrap can give perfect inference if the disturbances are symmetrically distributed about the origin.

Since exact results in bootstrap theory are very rare, and applicable only in very restricted circumstances, it is not surprising to find that, in general, the version of the wild bootstrap that gives perfect inference in one very restrictive case suffers from some size distortion. It appears, however, that the distortion is never more than that of any other version, as we demonstrate in a series of simulation experiments.

For these experiments, our policy is to concentrate on cases in which the asymptotic tests based on the HCCME are very badly behaved, and to try to identify bootstrap procedures that go furthest in correcting this bad behaviour. Thus, except for the purposes of obtaining benchmarks, we look at small samples of size 10, with an observation of very high leverage, and a great deal of heteroskedasticity closely correlated with the regressors.

It is of course important to study what happens when the disturbances are not symmetrically distributed. The asymptotic refinements found by Wu and Mammen for certain versions of the wild bootstrap are due to taking account of such skewness. We show the extent of the degradation in performance with asymmetric disturbances, but show that our preferred version of the wild bootstrap continues to work at least as well as any other, including the popular version of Liu and Mammen which takes explicit account of skewness.

Some readers may think it odd that we do not, in this paper, provide arguments based on Edgeworth expansions to justify and account for the phenomena that we illustrate by simulation. The reason is that the wild bootstrap uses a form of resampling, as a result of which the bootstrap distribution, conditional on the original data, is discrete. Since a discrete distribution cannot satisfy the Cramér condition, no valid Edgeworth expansion past terms of order $n^{-1/2}$ (n is the sample size) can exist for the distribution of wild bootstrap statistics; see for instance Kolassa (1994), Chapter 3, Bhattacharya and Ghosh (1978) Theorem 2, and Feller (1971), section XVI.4. Nonetheless, in Davidson and Flachaire (2001), we provide purely formal Edgeworth expansions that give heuristic support to the conclusions given here on the basis of simulation results.

In Section 2, we discuss a number of ways in which the wild bootstrap may be implemented, and show that, with symmetrically distributed disturbances, a property of independence holds that gives rise to an exact result concerning bootstrap P values. In Section 3, simulation experiments are described designed to measure the reliability of various tests, bootstrap and asymptotic, in various conditions, including very small samples, and to compare the rejection probabilities of these tests. These experiments give strong evidence in favour of our preferred version of the wild bootstrap. A few conclusions are drawn in Section 4.

2. The Wild Bootstrap

Consider the linear regression model

$$y_t = x_{t1}\beta_1 + \boldsymbol{X}_{t2}\boldsymbol{\beta}_2 + u_t, \quad t = 1, \ldots, n, \tag{1}$$

in which the explanatory variables are assumed to be strictly exogenous, in the sense that, for all t, x_{t1} and $\boldsymbol{X}_{t2}$ are independent of all of the disturbances u_s, $s = 1, \ldots, n$. The row vector $\boldsymbol{X}_{t2}$ contains observations on $k-1$ variables, of which, if $k > 1$, one is a constant. We wish to test the null hypothesis that the coefficient β_1 of the first regressor x_{t1} is zero.

The disturbances are assumed to be mutually independent and to have a common expectation of zero, but they may be heteroskedastic, with $\mathrm{E}(u_t^2) = \sigma_t^2$. We write $u_t = \sigma_t v_t$, where $\mathrm{E}(v_t^2) = 1$. We consider only *unconditional* heteroskedasticity, which means that the σ_t^2 may depend on the exogenous regressors, but not, for instance, on lagged dependent variables. The model represented by (1) is thus generated by the variation of the parameters β_1 and $\boldsymbol{\beta}_2$, the variances σ_t^2, and the probability distributions of the v_t. The regressors are taken as fixed and the same for all DGPs contained in the model. HCCME-based pseudo-t statistics for testing whether $\beta_1 = 0$ are then asymptotically pivotal for the restricted model in which we set $\beta_1 = 0$ if we also impose the weak condition that the σ_t^2 are bounded away from zero and infinity.

We write $\boldsymbol{x}_1$ for the n-vector with typical element x_{t1}, and $\boldsymbol{X}_2$ for the $n \times (k-1)$ matrix with typical row $\boldsymbol{X}_{t2}$. By $\boldsymbol{X}$ we mean the full $n \times k$ matrix $[\boldsymbol{x}_1 \;\; \boldsymbol{X}_2]$. Then the basic HCCME for the OLS parameter estimates of (1) is

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

where the $n \times n$ diagonal matrix $\hat{\boldsymbol{\Omega}}$ has typical diagonal element $\hat{u}_t^2$, where the $\hat{u}_t$ are the OLS residuals from the estimation either of the unconstrained model (1) or the constrained model in which $\beta_1 = 0$ is imposed. We refer to the version (2) of the HCCME as HC_0. Bias is reduced by multiplying the $\hat{u}_t$ by the square root of $n/(n-k)$, thereby multiplying the elements of $\hat{\boldsymbol{\Omega}}$ by $n/(n-k)$; this procedure, analogous to the use in the homoskedastic case of the unbiased OLS estimator of the variance of the disturbances, gives rise to form HC_1 of the HCCME. In the homoskedastic case, the variance of $\hat{u}_t$ is proportional to $1 - h_t$, where $h_t \equiv \boldsymbol{X}_t(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_t^\top$, the t^{th} diagonal element of the orthogonal projection matrix on to the span of the columns of $\boldsymbol{X}$. This suggests replacing the $\hat{u}_t$ by $\hat{u}_t/(1-h_t)^{1/2}$ in order to obtain $\hat{\boldsymbol{\Omega}}$. If this is done, we obtain form HC_2 of the HCCME. Finally, arguments based on the jackknife lead MacKinnon and White (1985) to propose form HC_3, for which the $\hat{u}_t$ are replaced by $\hat{u}_t/(1-h_t)$. MacKinnon and White and also Chesher and Jewitt (1987) show that, in terms of size distortion, HC_0 is outperformed by HC_1, which is in turn outperformed by 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.

As mentioned in the Introduction, heteroskedasticity of unknown form cannot be mimicked in the bootstrap distribution. The wild bootstrap gets round this

problem by using a bootstrap DGP of the form

$$y_t^* = \boldsymbol{X}_t \hat{\boldsymbol{\beta}} + u_t^*, \tag{3}$$

where $\hat{\boldsymbol{\beta}}$ is a vector of parameter estimates, and the bootstrap disturbance terms are

$$u_t^* = f_t(\hat{u}_t)\varepsilon_t, \tag{4}$$

where $f_t(\hat{u}_t)$ is a transformation of the OLS residual $\hat{u}_t$, and the ε_t are mutually independent drawings, completely independent of the original data, from some auxiliary distribution such that

$$\mathrm{E}(\varepsilon_t) = 0 \qquad \text{and} \qquad \mathrm{E}(\varepsilon_t^2) = 1. \tag{5}$$

Thus, for each bootstrap sample, the exogenous explanatory variables are reused unchanged, as are the OLS residuals $\hat{u}_t$ from the estimation using the original observed data. The transformation $f_t(\cdot)$ can be used to modify the residuals, for instance by dividing by $1 - h_t$, just as in the different variants of the HCCME.

In the literature, the further condition that $\mathrm{E}(\varepsilon_t^3) = 1$ is often added. Liu (1988) considers model (1) 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) suggested what is probably the most popular choice for the distribution of the ε_t, namely the following two-point distribution:

$$F_1: \quad \varepsilon_t = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } p = (\sqrt{5}+1)/(2\sqrt{5}) \\ (\sqrt{5}+1)/2 & \text{with probability } 1-p. \end{cases} \tag{6}$$

Liu also mentions the possibility of Rademacher variables, defined as

$$F_2: \quad \varepsilon_t = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2. \end{cases} \tag{7}$$

This distribution, for estimation of an expectation, satisfies necessary conditions for refinements in the case of unskewed disturbances. Unfortunately, she does not follow up this possibility, since (7), being a lattice distribution, does not lend itself to rigorous techniques based on Edgeworth expansion. In this paper, we show by other methods that (7) is, for all the cases we consider, the best choice of distribution for the ε_t. Another variant of the wild bootstrap that we consider later is obtained by replacing (4) by

$$u_t^* = f_t(|\hat{u}_t|)\varepsilon_t, \tag{8}$$

in which the absolute values of the residuals are used instead of the signed residuals.

Conditional on the random elements $\hat{\boldsymbol{\beta}}$ and $\hat{u}_t$, the wild bootstrap DGP (3) clearly belongs to the null hypothesis if the first component of $\hat{\boldsymbol{\beta}}$, corresponding to the regressor $\boldsymbol{x}_1$, is zero, since the bootstrap disturbance terms u_t^* have expectation zero and are heteroskedastic, for both formulations, (4) or (8), for any distribution

for the ε_t satisfying (5). Since (1) is linear, we may also set the remaining components of $\hat{\boldsymbol{\beta}}$ to zero, since the distribution of any HCCME-based pseudo-t statistic does not depend on the value of $\boldsymbol{\beta}_2$. Since the HCCME-based statistics we have discussed are asymptotically pivotal, inference based on the wild bootstrap using such a statistic applied to model (1) is asymptotically valid. In the case of a nonlinear regression, the distribution of the test statistic does depend on the specific value of $\boldsymbol{\beta}_2$, and so a consistent estimator of these parameters should be used in formulating the bootstrap DGP.

The arguments in Beran (1988) show that bootstrap inference benefits from asymptotic refinements if the random elements in the bootstrap DGP are consistent estimators of the corresponding elements in the unknown true DGP. These arguments do not apply directly to (3), since the squared residuals are not consistent estimators of the σ_t^2. In a somewhat different context from the present one, Davidson and MacKinnon (1999) show that bootstrap inference can be refined, sometimes beyond Beran's refinement, if the statistic that is bootstrapped is asymptotically independent of the bootstrap DGP. It is tempting to see if a similar refinement is available for the wild bootstrap. In what follows, we show that this is the case in some circumstances if the wild bootstrap makes use of the F_2 distribution, and that, in a very specific case, it leads to exact inference.

As discussed by Davidson and MacKinnon (1999), it is often useful for achieving this asymptotic independence to base the bootstrap DGP μ^* exclusively on estimates *under the null hypothesis*. Here, that means that we should set $\beta_1 = 0$ in the bootstrap DGP (3). Since we may also set $\boldsymbol{\beta}_2$ to zero, as remarked above, the bootstrap DGP can be expressed as

$$y_t^* = u_t^*, \qquad u_t^* = f_t(\tilde{u}_t)\varepsilon_t, \tag{9}$$

where the OLS residuals $\tilde{u}_t$ are obtained from the regression $y_t = \boldsymbol{X}_{t2}\boldsymbol{\beta}_2 + u_t$ that incorporates the constraint of the null hypothesis. In the case of nonlinear regression, the bootstrap DGP should be constructed using the NLS estimate $\tilde{\boldsymbol{\beta}}_2$ obtained by estimating the restricted model with $\beta_1 = 0$.

There is an important special case in which the wild bootstrap using F_2 yields almost perfect inference. This case arises when the entire parameter vector $\boldsymbol{\beta}$ vanishes under the null hypothesis and constrained residuals are used for both the HCCME and the wild bootstrap DGP. For convenience, we study bootstrap inference by looking at the bootstrap P value, defined as the probability mass in the bootstrap distribution in the region more extreme than the realised statistic.

Theorem 1

Consider the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t \tag{10}$$

where the $n \times k$ matrix $\boldsymbol{X}$ with typical row $\boldsymbol{X}_t$ is independent of all the disturbances u_t, assumed to have expectation zero, and to follow distributions symmetric about 0. Under the null hypothesis that $\boldsymbol{\beta} = \mathbf{0}$,

the χ^2 statistic for a test of that null against the alternative represented by (10), based on any of the four HCCMEs considered here constructed with constrained residuals, has exactly the same distribution as the same statistic bootstrapped, if the bootstrap DGP is the wild bootstrap (9), with $f(u) = u$ or equivalently $f(u) = |u|$, for which the ε_t are generated by the symmetric two-point distribution F_2 of (7).

For sample size n, the bootstrap P value p^* follows a discrete distribution supported by the set of points $p_i = i/2^n$, $i = 0, \ldots, 2^n - 1$, with equal probability mass 2^{-n} on each point. For each nominal level α equal to one of the dyadic numbers $1/2^n$, $i = 0, 1, \ldots, 2^n - 1$, the probability under the null hypothesis that the bootstrap P value is less than α, that is, the probability of Type I error at level α, is exactly equal to α

Proof:

The OLS estimates from (10) are given by $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$, and any of the HCCMEs we consider for $\hat{\boldsymbol{\beta}}$ can be written in the form (2), with an appropriate choice of $\hat{\boldsymbol{\Omega}}$. The χ^2 statistic thus takes the form

$$\tau \equiv \boldsymbol{y}^\top \boldsymbol{X} (\boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{11}$$

Under the null, $\boldsymbol{y} = \boldsymbol{u}$, and each component u_t of $\boldsymbol{u}$ can be written as $|u_t| s_t$, where s_t, equal to ± 1, is the sign of u_t, and is independent of $|u_t|$ because we assume that u_t follows a symmetric distribution. Define the $1 \times k$ row vector $\boldsymbol{Z}_t$ as $|u_t| \boldsymbol{X}_t$, and the $n \times 1$ column vector $\boldsymbol{s}$ with typical element s_t. The entire $n \times k$ matrix $\boldsymbol{Z}$ with typical row $\boldsymbol{Z}_t$ is then independent of the vector $\boldsymbol{s}$. If the constrained residuals, which are just the elements of $\boldsymbol{y}$, are used to form $\hat{\boldsymbol{\Omega}}$, the statistic (11) is equal to

$$\boldsymbol{s}^\top \boldsymbol{Z} \left(\sum_{t=1}^{n} a_t \boldsymbol{Z}_t^\top \boldsymbol{Z}_t \right)^{-1} \boldsymbol{Z}^\top \boldsymbol{s}, \tag{12}$$

where a_t is equal to 1 for HC_0, $n/(n-k)$ for HC_1, $1/(1-h_t)$ for HC_2, and $1/(1-h_t)^2$ for HC_3.

If we denote by τ^* the statistic generated by the wild bootstrap with F_2, then τ^* can be written as

$$\boldsymbol{\varepsilon}^\top \boldsymbol{Z} \left(\sum_{t=1}^{n} a_t \boldsymbol{Z}_t^\top \boldsymbol{Z}_t \right)^{-1} \boldsymbol{Z}^\top \boldsymbol{\varepsilon}, \tag{13}$$

where $\boldsymbol{\varepsilon}$ denotes the vector containing the ε_t. The matrix $\boldsymbol{Z}$ is exactly the same as in (12), because the exogenous matrix $\boldsymbol{X}$ is reused unchanged by the wild bootstrap, and the wild bootstrap disturbance terms $u_t^* = \pm u_t$, since, under F_2, $\varepsilon_t = \pm 1$. Thus, for all t, $|u_t^*| = |u_t|$. By construction, $\boldsymbol{\varepsilon}$ and $\boldsymbol{Z}$ are independent under the wild bootstrap DGP. But it is clear that $\boldsymbol{s}$ follows exactly the same distribution as $\boldsymbol{\varepsilon}$, and so it follows that τ under the null and τ^* under the wild bootstrap DGP with F_2 have the same distribution. This proves the first assertion of the theorem.

Conditional on the $|u_t|$, this common distribution of τ and τ^* is of course a discrete distribution, since $\boldsymbol{\varepsilon}$ and $\boldsymbol{s}$ can take on only 2^n different, equally probable, values, with a choice of $+1$ or -1 for each of the n components of the vector. In fact, there are normally only 2^{n-1} different values, because, given that (13) is a quadratic form in $\boldsymbol{\varepsilon}$, the statistic for $-\boldsymbol{\varepsilon}$ is the same as for $\boldsymbol{\varepsilon}$. However, if there is only one degree of freedom, one may take the signed square root of (13), in which case the symmetry is broken, and the number of possible values is again equal to 2^n. For the rest of this proof, therefore, we consider the case with 2^n possibilities.

The statistic τ must take on one of these 2^n possible values, each with the same probability of 2^{-n}. If we denote the 2^n values, arranged in increasing order, as τ_i, $i = 1, \dots, 2^n$, with $\tau_j > \tau_i$ for $j > i$, then, if $\tau = \tau_i$, the bootstrap P value, which is the probability mass in the distribution to the right of τ_i, is just $1 - i/2^n$. As i ranges from 1 to 2^n, the P value varies over the set of points $p_i \equiv i/2^n$, $i = 0, \dots, 2^n - 1$, all with probability 2^{-n}. This distribution, conditional on the $|u_t|$, does not depend on the $|u_t|$, and so is also the unconditional distribution of the bootstrap P value.

For nominal level α, the bootstrap test rejects if the bootstrap P value is less than α. To compute this probability, we consider the rank i of the realised statistic τ in the set of 2^n possibilities as a random variable uniformly distributed over the values $1, \dots, 2^n$. The bootstrap P value, equal to $1 - i/2^n$, is less than α if and only if $i > 2^n(1 - \alpha)$, an event of which the probability is 2^{-n} times the number of integers in the range $\lfloor 2^n(1 - \alpha) \rfloor + 1, \dots 2^n$, where $\lfloor x \rfloor$ denotes the greatest integer not greater than x. Let the integer k be equal to $\lfloor 2^n(1 - \alpha) \rfloor$. Then the probability we wish to compute is $2^{-n}(2^n - k) = 1 - k2^{-n}$.

Suppose first that this probability is equal to α. Then $\alpha 2^n$ is necessarily an integer, so that α is one of the dyadic numbers mentioned in the statement of the theorem. Suppose next that $\alpha = j/2^n$, j an integer. Then $\lfloor 2^n(1 - \alpha) \rfloor = 2^n - j$, so that $k = 2^n - j$. The probability of rejection by the bootstrap test at level α is therefore $1 - k2^{-n} = j/2^n = \alpha$. This proves the final assertions of the Theorem. ■

Remarks: For small enough n, it may be quite feasible to enumerate all the possible values of the bootstrap statistic τ^*, and thus obtain the exact value of the realisation p^*.

Although the discrete nature of the bootstrap distribution means that it is not possible to perform exact inference for an arbitrary significance level α, the problem is no different from the problem of inference with any discrete-valued statistic. For the case with $n = 10$, which will be extensively treated in the following sections, $2^n = 1024$, and so the bootstrap P value cannot be in error by more than 1 part in a thousand.

It is possible to imagine a case in which the discreteness problem is aggravated by the coincidence of some adjacent values of the τ_i of the proof of the theorem. For instance, if the only regressor in $\boldsymbol{X}$ is the constant, the value of (12) depends only on the number of positive components of $\boldsymbol{s}$ and not on their ordering. For this case, of course, it is not necessary to base inference on an HCCME. Coincidence of values of the τ_i will otherwise occur if all the explanatory variables take on exactly the same values for more than one observation. However, since this phenomenon

is observable, it need not be a cause for concern. A very small change in the values of the components of the $\boldsymbol{X}_t$ would be enough to break the ties in the τ_i.

The exact result of the theorem is specific to the wild bootstrap with F_2. The proof works because the signs in the vector $\boldsymbol{s}$ also follow the distribution F_2. Given the exact result of the theorem, it is of great interest to see the extent of the size distortion of the F_2 bootstrap with constrained residuals when the null hypothesis involves only a subset of the regression parameters. This question will be investigated by simulation in the following section. At this stage, it is possible to see why the theorem does not apply more generally. The expressions (12) and (13) for τ and τ^* continue to hold if the constrained residuals $\tilde{u}_t$ are used for $\hat{\boldsymbol{\Omega}}$, and if $\boldsymbol{Z}_t$ is redefined as $|\tilde{u}_t|(\boldsymbol{M}_2\boldsymbol{X}_1)_t$, where $\boldsymbol{X}_1$ is the matrix of regressors admitted only under the alternative, and $\boldsymbol{M}_2$ is the projection off the space spanned by the regressors that are present under the null. However, although $\boldsymbol{\varepsilon}$ in τ^* is by construction independent of $\boldsymbol{Z}$, $\boldsymbol{s}$ in τ is not. This is because the covariance matrix of the residual vector $\tilde{\boldsymbol{u}}$ is not diagonal in general, unlike that of the disturbances $\boldsymbol{u}$. In Figure 1, this point is illustrated for the bivariate case. In panel a), two level curves are shown of the joint density of two symmetrically distributed and independent variables u_1 and u_2. In panel b), the two variables are no longer independent. For the set of four points for which the absolute values of u_1 and u_2 are the same, it can be seen that, with independence, all four points lie on the same level curve of the joint density, but that this is no longer true without independence. The *vector* of absolute values is no longer independent of the *vector* of signs, even though independence still holds for the marginal distribution of each variable. Of course the asymptotic distributions of τ and τ^* still coincide.

3. Experimental Design and Simulation Results

It was shown by Chesher and Jewitt (1987) that HCCMEs are most severely biased when the regression design has observations with high leverage, and that the extent of the bias depends on the amount of heteroskedasticity in the true DGP. Since in addition one expects bootstrap tests to behave better in large samples than in small, in order to stress-test the wild bootstrap, most of our experiments are performed with a sample of size 10 containing one regressor, denoted $\boldsymbol{x}_1$, all the elements but one of which are independent drawings from $\mathrm{N}(0,1)$, but the second of which is 10, so as to create an observation with exceedingly high leverage. All the tests we consider are of the null hypothesis that $\beta_1 = 0$ in the model (1), with k, the total number of regressors, varying across experiments. In all regression designs, $\boldsymbol{x}_1$ is always present; for the design we designate by $k = 2$ a constant, denoted $\boldsymbol{x}_2$, is also present; and for $k = 3, \ldots, 6$, additional regressors $\boldsymbol{x}_i$, $i = 3, \ldots, 6$ are successively appended to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. In Table 1, the components of $\boldsymbol{x}_1$ are given, along with those of the $\boldsymbol{x}_i$, $i = 3, \ldots, 6$. In Table 2 are given the diagonal elements h_t of the orthogonal projections on to spaces spanned by $\boldsymbol{x}_1$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$, *etc.* The h_t measure the leverage of the 10 observations for the different regression designs.

The data in all the simulation experiments discussed here are generated under the null hypothesis. Since (1) is a linear model, we set $\boldsymbol{\beta}_2 = \mathbf{0}$ without loss of

generality. Thus our data are generated by a DGP of the form

$$y_t = \sigma_t v_t, \qquad t = 1, \ldots, n, \tag{14}$$

where n is the sample size, 10 for most experiments. For homoskedastic data, we set $\sigma_t = 1$ for all t, and for heteroskedastic data, we set $\sigma_t = |x_{t1}|$, the absolute value of the t^{th} component of $\boldsymbol{x}_1$. Because of the high leverage observation, this gives rise to very strong heteroskedasticity, which leads to serious bias of the OLS covariance matrix; see White (1980). The v_t are independent variables of zero expectation and unit variance, and in the experiments will be either normal or else drawings from the highly skewed $\chi^2(2)$ distribution, centred and standardised.

The main object of our experiments is to compare the size distortions of wild bootstrap tests using the distributions F_1 and F_2. Although the latter gives exact inference only in a very restricted case, we show that it always leads to less distortion than the former in sample sizes up to 100. We also conduct a few experiments comparing the wild bootstrap and the (y, X) bootstrap. In order to conduct a fair comparison, we use an improved version of the (y, X) bootstrap suggested by Mammen (1993), and subsequently modified by Flachaire (1999), in which we resample, not the (y, X) pairs as such, but rather regressors (X) and the constrained residuals, transformed according to HC_3. For the wild bootstrap, we are also interested in the impact on errors in the rejection probabilities (ERPs) of the use of unconstrained versus constrained residuals, and the use of the different sorts of HCCME. Here, we formally define the error in rejection probability as the difference between the rejection probability, as estimated by simulation, at a given nominal level, and that nominal level.

We present our results as P value discrepancy plots, as described in Davidson and MacKinnon (1998). These plots show ERPs as a function of the nominal level α. Since we are considering a one-degree-of-freedom test, it is possible to perform a one-tailed test for which the rejection region is the set of values of the statistic algebraically greater than the critical value. We choose to look at one-tailed tests because Edgeworth expansions predict – see Hall (1992) – that the ERPs of one-tailed bootstrap tests converge to zero with increasing sample size more slowly than those of two-tailed tests. In any event, it is easy to compute the ERP of a two-tailed test with the information in the P value discrepancy plot. All plots are based on experiments using $100,000$ replications.

We now present our results as answers to a series of pertinent questions.

- In a representative case, with strong heteroskedasticity and high leverage, is the wild bootstrap capable of reducing the ERP relative to asymptotic tests?

Figure 2 shows plots for the regression design with $k = 3$, sample size $n = 10$, and normal heteroskedastic disturbances. The ERPs are plotted for the conventional t statistic, based on the OLS covariance matrix estimate, the four versions of HCCME-based statistics, HC_i, $i = 0, 1, 2, 3$, all using constrained residuals. P values for the asymptotic tests are obtained using Student's t distribution with 7 degrees of freedom. The ERP is also plotted for what will serve as a base case for the wild bootstrap: Constrained residuals are used both for the HCCME and the wild bootstrap DGP, the F_2 distribution is used for the ε_t, and the statistic that

is bootstrapped is the HC_3 form. To avoid redundancy, the plots are drawn only for the range $0 \leq \alpha \leq 0.5$, since all these statistics are symmetrically distributed when the disturbances are symmetric. In addition, the bootstrap statistics are symmetrically distributed conditional on the original data, and so the distribution of the bootstrap P value is also symmetrical about $\alpha = 0.5$. It follows that the ERP for nominal level α is the negative of that for $1 - \alpha$. Not surprisingly, the conventional t statistic, which does not have even an asymptotic justification, is the worst behaved of all, with far too much mass in the tails. But, although the HC_i statistics are less distorted, the bootstrap test is manifestly much better behaved.

- The design with $k = 1$ satisfies the conditions of Theorem 1 when the disturbances are symmetric and the HCCME and the bootstrap DGP are based on constrained residuals. If we maintain all these conditions but consider the cases with $k > 1$, bootstrap inference is no longer perfect. To what extent is inference degraded?

P value discrepancy plots are shown in Figure 3 for the designs $k = 1, \ldots, 6$ using the base-case wild bootstrap as described above. Disturbances are normal and heteroskedastic. As expected, the ERP for $k = 1$ is just experimental noise, and for most other cases the ERPs are significant. By what is presumably a coincidence induced by the specific form of the data, they are not at all large for $k = 5$ or $k = 6$. In any case, we can conclude that the ERP does indeed depend on the regression design, but quantitatively not very much, given the small sample size.

- How do bootstrap tests based on the F_1 and F_2 distributions compare? We expect that F_2 will lead to smaller ERPs if the disturbances are symmetric, but what if they are asymmetric? How effective is the skewness correction provided by F_1? What about the (y, X) bootstrap?

In Figure 4 plots are shown for the $k = 3$ design with heteroskedastic normal disturbances and skewed $\chi^2(2)$ disturbances. The F_1 and F_2 bootstraps give rather similar ERPs, whether or not the disturbances are skewed. But the F_2 bootstrap is generally better, and never worse. Very similar results, leading to same conclusion, were also obtained with the $k = 4$ design. For $k = 1$ and $k = 2$, on the other hand, the F_1 bootstrap suffers from larger ERPs than for $k > 2$. Plots are also shown for the same designs and the preferred form of the (y, X) bootstrap. It is clear that the ERPs are quite different from those of the wild bootstrap, in either of its forms, and substantially greater.

- What is the penalty for using the wild bootstrap when the disturbances are homoskedastic and inference based on the conventional t statistic is reliable, at least with normal disturbances? Do we get different answers for F_1, F_2, and the (y, X) bootstrap?

Again we use the $k = 3$ design. We see from Figure 5, which is like Figure 4 except that the disturbances are homoskedastic, that, with normal disturbances, the ERP is very slight with F_2, but remains significant for F_1 and (y, X). Thus, with unskewed, homoskedastic disturbances, the penalty attached to using the F_2 bootstrap is very small. With skewed disturbances, all three tests give substantially greater ERPs, but the F_2 version remains a good deal better than the F_1 version, which in turn is somewhat better than the (y, X) bootstrap.

- Do the rankings of bootstrap procedures obtained so far for $n = 10$ continue to apply for larger samples? Do the ERPs become smaller rapidly as n grows?

In order to deal with larger samples, the data in Table 1 were simply repeated as needed in order to generate regressors for $n = 20, 30, \ldots$. The plots shown in Figures 4 and 5 are repeated in Figure 6 for $n = 100$. The rankings found for $n = 10$ remain unchanged, but the ERP for the F_2 bootstrap with skewed, heteroskedastic, disturbances improves less than that for the F_1 bootstrap with the increase in sample size. It is noteworthy that none of the ERPs in this diagram is very large.

In Figure 7, we plot the ERP for $\alpha = 0.05$ as a function of n, $n = 10, 20, \ldots$, with the $k = 3$ design and heteroskedastic disturbances, normal for F_1 and skewed for F_2, chosen because these configurations lead to comparable ERPs for n around 100, and because this is the worst setup for the F_2 bootstrap. It is interesting to observe that, at least for $\alpha = 0.05$, the ERPs are not monotonic. What seems clear is that, although the absolute magnitude of the ERPs is not disturbingly great, the rate of convergence to zero does not seem to be at all rapid, and seems to be slower for the F_2 bootstrap.

We now move on to consider some lesser questions, the answers to which justify, at least partially, the choices made in the design of our earlier experiments. We restrict attention to the F_2 bootstrap, since it is clearly the procedure of choice in practice.

- Does it matter which of the four versions of the HCCME is used?

It is clear from Figure 2 that the choice of HC_i has a substantial impact on the ERP of the asymptotic test. Since the HC_0 and HC_1 statistics differ only by a constant multiplicative factor, they yield identical bootstrap P values, as do all versions for $k = 1$ and $k = 2$. For $k = 1$ this is obvious, since the raw statistics are identical, and for $k = 2$, the only regressor other than $\boldsymbol{x}_1$ is the constant, and so h_t does not depend on t. For $k > 2$, significant differences appear, as seen in Figure 8 which treats the $k = 4$ design. HC_3 has the least distortion here, and also for the other designs with $k > 2$. This accounts for our choice of HC_3 in the base case.

- What is the best transformation $f_t(\cdot)$ to use in the definition of the bootstrap DGP? Plausible answers are either the identity transformation, or the same as that used for the HCCME.

No very clear answer to this question emerged from our numerous experiments on this point. A slight tendency in favour of using the HC_3 transformation appears, but this choice does not lead to universally smaller ERPs. However, the quantitative impact of the choice is never very large, and so the HC_3 transformation is used in our base case.

- How is performance affected if the leverage of observation 2 is reduced?

Because the ERPs of the asymptotic tests are greater with a high leverage observation, we might expect the same to be true of bootstrap tests. In fact, although this is true if the HC_0 statistic is used, the use of widely varying h_t with HC_3 provides a good enough correction that, with it, the presence or absence of leverage has little impact. In Figure 9, this is demonstrated for $k = 3$, and normal disturbances,

and the effect of leverage is compared with that of heteroskedasticity. The latter is clearly a much more important determinant of the ERP than the former. Similar results are obtained if the null hypothesis concerns a coefficient other than β_1. In that case, the h_t differ more among themselves, since $\boldsymbol{x}_1$ is now used in their calculation, and HC_0 gives more variable results than HC_3, for which the ERPs are similar in magnitude to those for the test of $\beta_1 = 0$.

- How important is it to use constrained residuals?

For Theorem 1 to hold, it is essential. Simulation results show that, except for the $k = 1$ and $k = 2$ designs, it is not very important whether one uses constrained or unconstrained residuals, although results with constrained residuals tend to be better in most cases. The simulations do however show clearly that it is a mistake to mix unconstrained residuals in the HCCME and constrained residuals for the bootstrap DGP.

4. Conclusion

The wild bootstrap is commonly applied to models with heteroskedastic disturbances and an unknown pattern of heteroskedasticity, most commonly in the form that uses the asymmetric F_1 distribution in order to take account of skewness of the disturbances. In this paper we have shown that the wild bootstrap implemented with the symmetric F_2 distribution and constrained residuals, which can give perfect inference in one very restricted case, is never any worse behaved than the F_1 version, or either version with unconstrained residuals, and is usually markedly better. We therefore recommend that this version of the wild bootstrap should be used in practice in preference to other versions. This recommendation is supported by the results of simulation experiments designed to expose potential weaknesses of both versions.

It is important to note that conventional confidence intervals cannot benefit from our recommended version of the wild bootstrap, since they are implicitly based on a Wald test using unconstrained residuals for the HCCME and, unless special precautions are taken, also for the bootstrap DGP. This is not a problem for many econometric applications, for which hypothesis tests may be sufficient. In those cases in which reliable confidence intervals are essential, we recommend that they be obtained by inverting a set of tests based on the preferred wild bootstrap. Although this can be a computationally intensive procedure, it is well within the capacity of modern computers and seems to be the only way currently known to extend refinements available for tests to confidence intervals.

A final caveat seems called for: Although our experiments cover a good number of cases, some caution is still necessary on account of the fact that the extent of the ERP of wild bootstrap tests appears to be sensitive to details of the regression design and the pattern of heteroskedasticity.

In this paper, we have tried to investigate worst case scenarios for wild bootstrap tests. This should not lead readers to conclude that the wild bootstrap is an unreliable method in practice. On the contrary, as Figure 6 makes clear, it suffers from very little distortion for samples of moderate size unless there is extreme

heteroskedasticity. In most practical contexts, use of the F_2-based wild bootstrap with constrained residuals should provide satisfactory inference.

References

Beran, R. (1986). Discussion of "Jackknife bootstrap and other resampling methods in regression analysis" by C. F. J. Wu., *Annals of Statistics* 14, 1295–1298.

Beran, R. (1988). "Prepivoting test statistics: a bootstrap view of asymptotic refinements", *Journal of the American Statistical Association*, 83, 687–697.

Bhattacharya, R. N., and J. K. Ghosh (1978). "On the validity of the formal Edgeworth expansion", *Annals of Statistics*, 6, 434–451.

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

Davidson, R. and E. Flachaire (2001). "The Wild Bootstrap, Tamed at Last", Queen's Economics Department working paper #1000, Queen's University, Kingston, Canada.

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

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

Eicker, F. (1963). "Asymptotic normality and consistency of the least squares estimators for families of linear regressions", *The Annals of Mathematical Statistics*, 34, 447–456.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, second edition, Wiley, New York.

Flachaire, E. (1999). "A better way to bootstrap pairs", *Economics Letters*, 64, 257–262.

Freedman, D. A. (1981). "Bootstrapping regression models", *Annals of Statistics*, 9, 1218–1228.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

Kolassa, J. E. (1994). *Series Approximations in Statistics*, Lecture Notes in Statistics 88, Springer-Verlag, New York.

Liu, R. Y. (1988). "Bootstrap procedures under some non-I.I.D. models", *Annals of Statistics* 16, 1696–1708.

MacKinnon, J. G., and H. White (1985). "Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties", *Journal of Econometrics*, 29, 305–325.

Mammen, E. (1993). "Bootstrap and wild bootstrap for high dimensional linear models", *Annals of Statistics* 21, 255–285.

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", *Annals of Statistics* 14, 1261–1295.

Table 1. Regressors

Obs	$\boldsymbol{x}_1$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$
1	0.616572	0.511730	0.210851	-0.651571	0.509960
2	10.000000	5.179612	4.749082	6.441719	1.212823
3	-0.600679	0.255896	-0.150372	-0.530344	0.318283
4	-0.613076	0.705476	0.447747	-1.599614	-0.601335
5	-1.972106	-0.673980	-1.513501	0.533987	0.654767
6	0.409741	0.922026	1.162060	-1.328799	1.607007
7	-0.676614	0.515275	-0.241203	-1.424305	-0.360405
8	0.400136	0.459530	0.166282	0.040292	-0.018642
9	1.106144	2.509302	0.899661	-0.188744	1.031873
10	0.671560	0.454057	-0.584329	1.451838	0.665312

Table 2. Leverage measures

Obs	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
1	0.003537	0.101022	0.166729	0.171154	0.520204	0.560430
2	0.930524	0.932384	0.938546	0.938546	0.964345	0.975830
3	0.003357	0.123858	0.128490	0.137478	0.164178	0.167921
4	0.003497	0.124245	0.167158	0.287375	0.302328	0.642507
5	0.036190	0.185542	0.244940	0.338273	0.734293	0.741480
6	0.001562	0.102785	0.105276	0.494926	0.506885	0.880235
7	0.004260	0.126277	0.138399	0.143264	0.295007	0.386285
8	0.001490	0.102888	0.154378	0.162269	0.163588	0.218167
9	0.011385	0.100300	0.761333	0.879942	0.880331	0.930175
10	0.004197	0.100698	0.194752	0.446773	0.468841	0.496971

Notes: For $k = 1$, the only regressor is $\boldsymbol{x}_1$, for $k = 2$ there is also the constant, for $k = 3$ there are the constant, $\boldsymbol{x}_1$, and $\boldsymbol{x}_2$, and so forth.

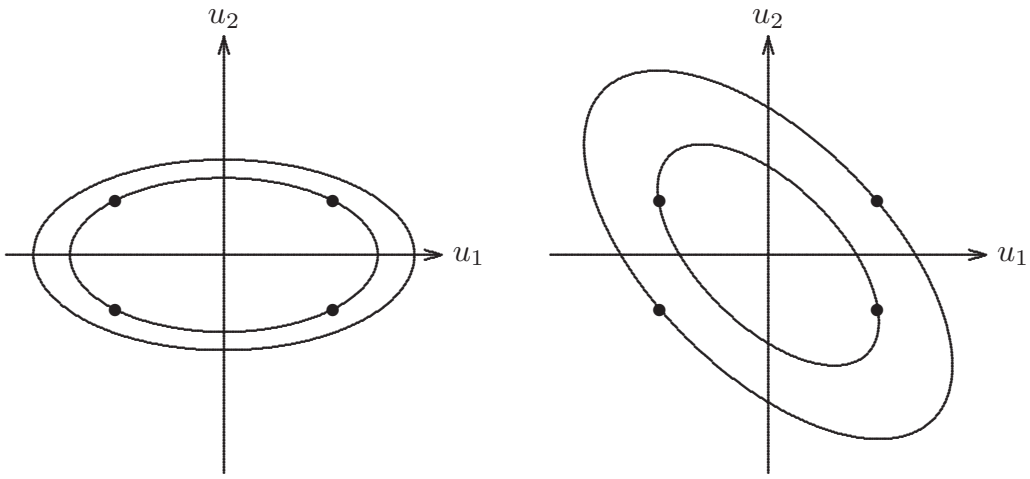

Figure 1. Absolute values and signs of two random variables

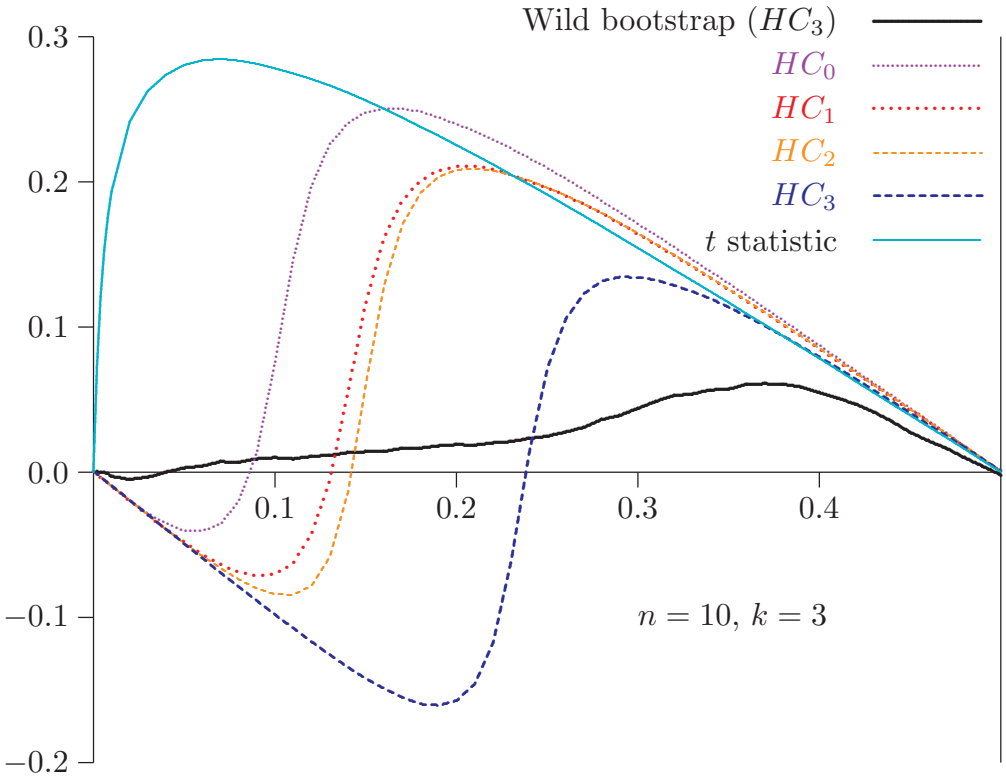

Figure 2. ERPs of asymptotic and bootstrap tests

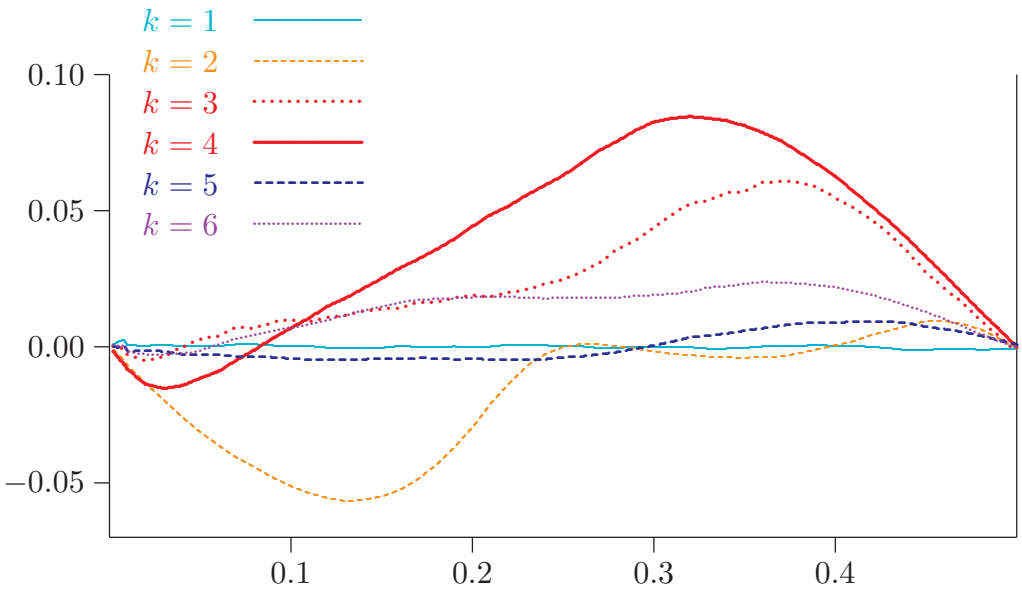

Figure 3. Base case with different designs

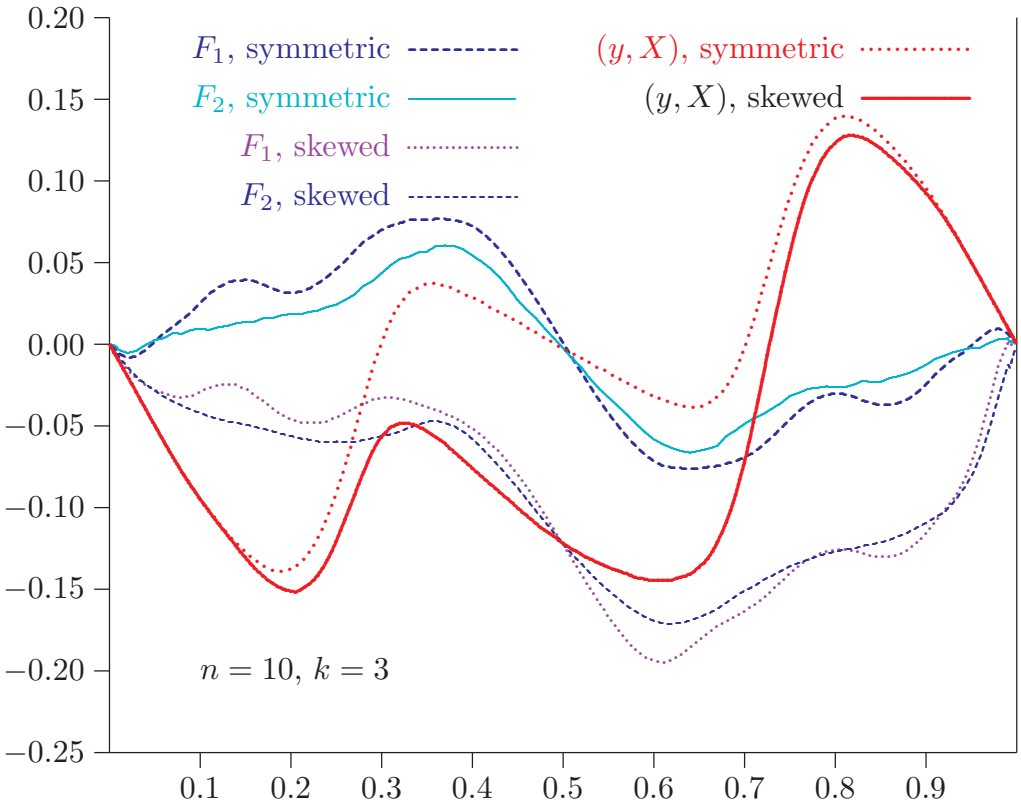

Figure 4. Symmetric and skewed errors, F_1, F_2, and (y, X) bootstraps

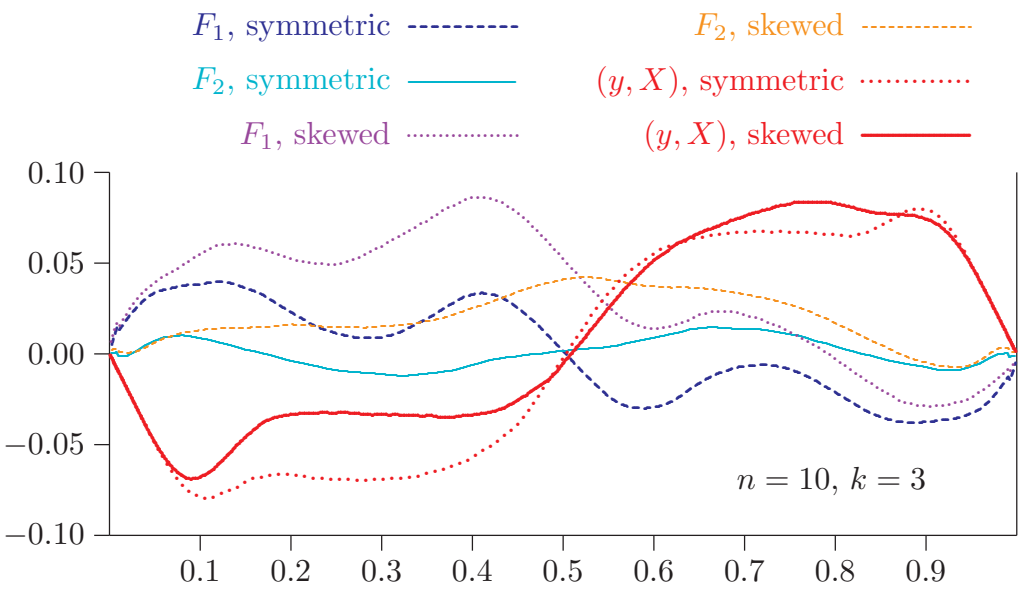

Figure 5. Homoskedastic errors

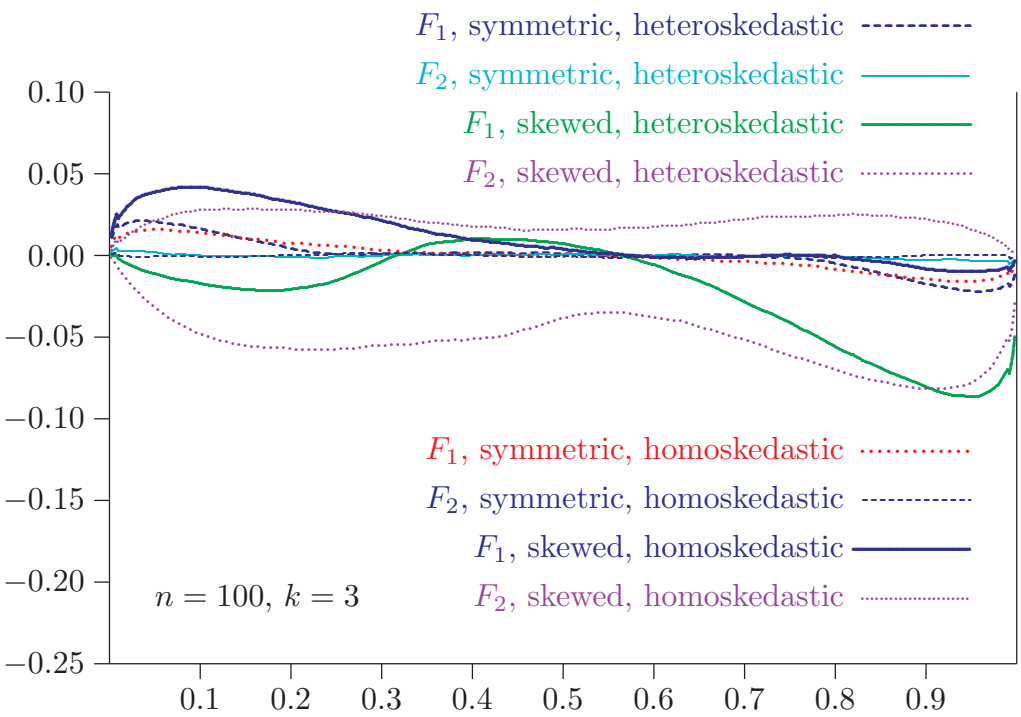

Figure 6. ERPs for $n = 100$

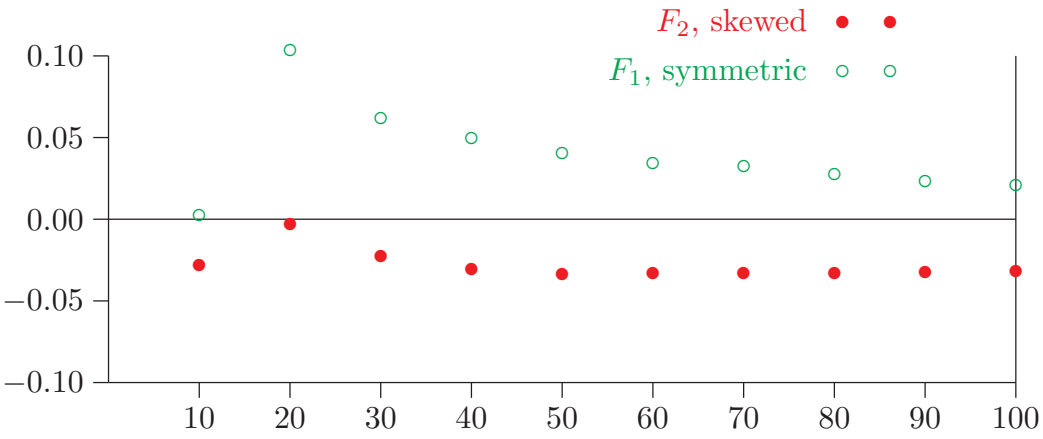

Figure 7. ERP for nominal level 0.05 as function of sample size

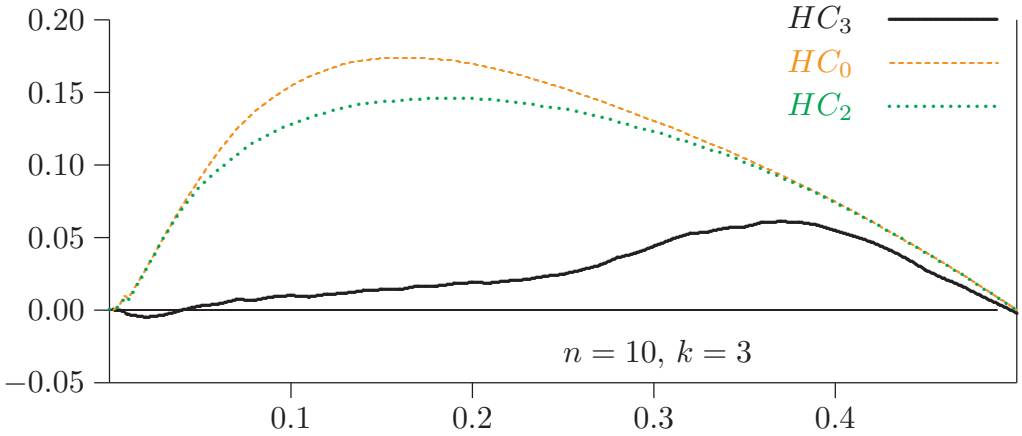

Figure 8. HC_3 compared with HC_0 and HC_2

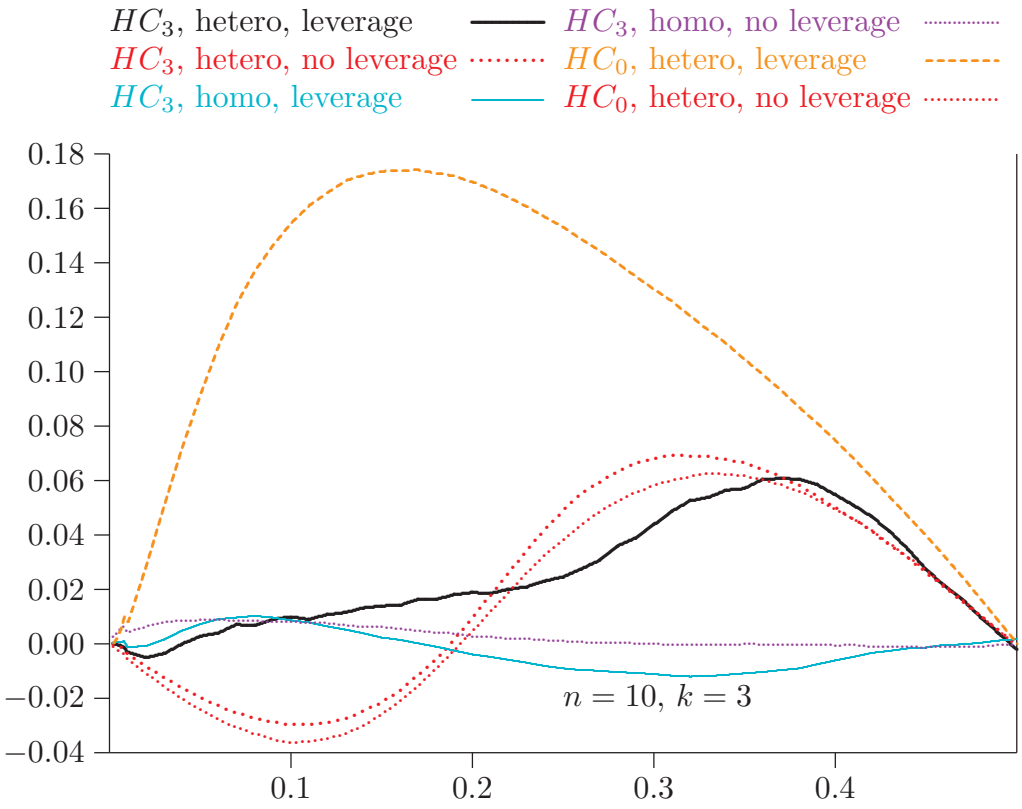

Figure 9. Relative importance of leverage and heteroskedasticity