The Wild Bootstrap, Tamed at Last

by

Russell Davidson

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

russell@ehess.cnrs-mrs.fr

and

Emmanuel Flachaire

CORE
34 voie du Roman Pays
1348 Louvain-la-Neuve, Belgium

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

flachaire@core.ucl.ac.be

Abstract

In this paper we are interested in inference based on heteroskedasticity consistent covariance matrix estimators, for which the appropriate bootstrap is a version of the wild bootstrap. Simulation results, obtained by a new very efficient method, show that all wild bootstrap tests exhibit substantial size distortion if the error terms are skewed and strongly heteroskedastic. The distortion is however less, sometimes much less, if one uses a version of the wild bootstrap, belonging to a class we call "tamed", which benefit from an asymptotic refinement related to the asymptotic independence of the bootstrapped test statistic and the bootstrap DGP. This version always gives better results than the version usually recommended in the literature, and gives exact results for some specific cases. However, when exact results are not available, we find that the rate of convergence to zero of the size distortion of wild bootstrap tests is not very rapid: in some cases, significant size distortion still remains for samples of size 100.

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, and to participants at the ESRC Econometrics Conference (Bristol), especially Whitney Newey. Remaining errors are ours.

July, 1999

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, $\boldsymbol{X}$ is 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 error terms $\boldsymbol{u}$ are heteroskedastic. In that case, the usual OLS estimator of the covariance of the OLS estimates $\hat{\boldsymbol{\beta}}$ is in general 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.

A technique that has been used to overcome this last difficulty is the so-called wild bootstrap. The 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 errors, and further evidence was provided by 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, which lead to agreement between the distributions of the raw and bootstrap statistics to higher than leading order asymptotically. However, neither Wu nor Mammen considered the case of HCCME-based statistics in a regression model with several regressors.

In this paper, we consider a number of implementations both of the Eicker-White HCCME and of the wild bootstrap applied to them. We show that, when the error

terms are symmetrically distributed about the origin, and when both the HCCME and the wild bootstrap DGP are based on residuals obtained by estimation under the null hypothesis, statistics based on all of the implementations that we study of the HCCME are asymptotically independent of the random elements that determine the wild bootstrap DGP. Davidson and MacKinnon (1999) have shown that such asymptotic independence leads to asymptotic refinements of bootstrap inference. We are able to go further when the hypothesis under test is that all the regression parameters are zero. In that event, we show that one version of the wild bootstrap gives essentially perfect inference.

In general, this version of the wild bootstrap suffers from some size distortion, but, it would appear, never more than 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. We use a method recently developed by Davidson and MacKinnon (1999), applicable in cases in which the test statistic and the bootstrap DGP are asymptotically independent, in order to perform efficient simulations that estimate the difference between the true and nominal rejection probabilities of bootstrap tests.

Another question of some importance is what happens when the error terms are not symmetrically distributed. The asymptotic refinements found by Wu and Mammen for certain versions of the wild bootstrap are directed at taking account of such skewness. However in this case, the asymptotic independence of the statistic and the bootstrap DGP no longer holds. In addition, we can no longer use the Davidson-MacKinnon trick for efficient estimation of the bootstrap size distortion. However it is possible, by a procedure we call *taming*, to restore asymptotic independence. Simulation experiments, including some more costly ones, in which a full bootstrap test is performed on each replication, show the extent of the degradation in performance with asymmetric error terms, but confirm that taming leads to somewhat better behaviour, and that our preferred version of the wild bootstrap, which is tamed by construction, continues to work at least as well as any other.

In section 2, we review the properties of bootstrap P values, and the circumstances in which they may benefit from refinements of various sorts. In section 3, we discuss a number of ways in which the wild bootstrap may be implemented, and show that, with symmetrically distributed error terms, a property of asymptotic independence holds that gives rise to an asymptotic refinement of bootstrap P values. In some special cases, the refinement can give rise to essentially exact inference. Section 4 reviews a method presented by Davidson and MacKinnon (1999) whereby the size distortion of bootstrap tests can be efficiently estimated by simulation without needing to do Monte Carlo on the bootstrap, and, in section 5, a series of experiments are described in which this method is used wherever possible. Section 6 contains the results of these experiments, and there are a few conclusions in section 7.

2. Bootstrap *P* Values

Beran (1988) showed that bootstrap inference is refined when the quantity bootstrapped is asymptotically pivotal. It is convenient to formalise the idea of pivotalness by means of a few formal definitions. A *data-generating process*, or *DGP*, is any rule sufficiently specific to allow artificial samples of arbitrary size to be simulated on the computer. Thus all parameter values and all probability distributions must be provided in the specification of a DGP. A *model* is a set of DGPs. Models are usually generated by allowing parameters and probability distributions to vary over admissible sets. A test statistic is a random variable that is a deterministic function of the data generated by a DGP and, possibly, other exogenous variables. A test statistic τ is a *pivot* for a model $\mathbb{M}$ if, for each sample size n, its distribution is independent of the DGP $\mu \in \mathbb{M}$ which generates the data from which τ is calculated. The *asymptotic distribution* of a test statistic τ for a DGP μ is the limit, if it exists, of the distribution of τ under μ as the sample size tends to infinity. The statistic τ is *asymptotically pivotal* for $\mathbb{M}$ if its asymptotic distribution exists for all $\mu \in \mathbb{M}$ and is independent of μ.

In hypothesis testing, the null hypothesis under test is represented by a model, as defined above. A test statistic is said to be pivotal or asymptotically pivotal under the null hypothesis if it is a pivot or an asymptotic pivot for the model that represents the hypothesis. Most test statistics commonly used in econometric practice are asymptotically pivotal under the null hypotheses they test, since asymptotically they have distributions, like standard normal, or chi-squared, that do not depend on unknown parameters. Conventional asymptotic inference is based on these known asymptotic distributions.

Even if a statistic is an exact pivot for a model $\mathbb{M}$, asymptotic inference may be only approximate, since the finite-sample distribution of the statistic may be the same for all $\mu \in \mathbb{M}$, but different for different sample sizes. A simple example is the use of an exact t statistic with asymptotic standard normal critical values. Bootstrap inference corrects this by relying on the finite-sample distribution of some specific DGP in the model $\mathbb{M}$ that represents the null hypothesis. Since for a pivot the choice of that DGP has no influence on the distribution, exact inference is possible. Usually, of course, the finite-sample distribution is estimated by simulation, with the consequent introduction of simulation error. Since this error can be made arbitrarily small by increasing the number of simulations, we will not concern ourselves with it here.

If an asymptotic pivot τ is not an exact pivot, its distribution depends on which particular DGP $\mu \in \mathbb{M}$ generates the data used to compute it. In this case, bootstrap inference is no longer exact in general. The bootstrap samples used to estimate the finite-sample distribution of τ are generated by a *bootstrap DGP*, which, although it usually belongs to $\mathbb{M}$, is in general different from the DGP that generated the original data.

It is possible to use the bootstrap either to calculate a critical value for τ or to calculate the marginal significance level, or P value, associated with a realisation of it. In this paper, we prefer the latter approach, as it greatly simplifies the analysis. Suppose that data are generated by a DGP μ_0 belonging to $\mathbb{M}$, and used

to compute a realisation $\hat{\tau}$ of the random variable τ. Then, for a test that rejects for large values of the statistic, the P value we would ideally like to compute is

$$p(\hat{\tau}) \equiv \Pr_{\mu_0}(\tau > \hat{\tau}). \tag{1}$$

In practice, (1) cannot be computed, or estimated by simulation, because the DGP μ_0 that generates observed data is unknown. If τ is an exact pivot, this does not matter, since (1) can be computed using any DGP in $\mathbb{M}$. In this case, $p(\hat{\tau})$ is a drawing from the $U(0,1)$ distribution. If τ is only an asymptotic pivot, the *bootstrap P value* is defined by

$$p^*(\hat{\tau}) \equiv \Pr_{\hat{\mu}}(\tau > \hat{\tau}), \tag{2}$$

where $\hat{\mu}$ is a (random) bootstrap DGP in $\mathbb{M}$, determined in some suitable way from the same data as those used to compute $\hat{\tau}$. We denote by $\hat{p}^*$ the bootstrap P value (2), and by p^* the random variable of which it is a realisation. Similarly, μ^* denotes the random DGP of which $\hat{\mu}$ is a realisation.

Let the asymptotic CDF of the asymptotic pivot τ be denoted by F. At nominal level α, an asymptotic test rejects if the asymptotic P value $1 - F(\hat{\tau}) < \alpha$. In order to avoid having to deal with different asymptotic distributions, it is convenient to replace the raw statistic τ by the asymptotic P value $1 - F(\tau)$, of which the asymptotic distribution is always $U(0,1)$. Henceforth, τ denotes such an asymptotic P value.

For the sample size of the observed data, the "rejection probability function," or RPF, provides a measure of the true rejection probability of the asymptotic test. This function, which gives the rejection probability under μ of a test at nominal level α, is defined as follows:

$$R(\alpha, \mu) \equiv \Pr_{\mu}\big(\tau < \alpha\big). \tag{3}$$

It is clear that $R(\cdot, \mu)$ is the CDF of τ under μ. The information contained in the function R is also provided by the "critical value function," or CVF, Q, defined implicitly by the equation

$$\Pr_{\mu}\big(\tau < Q(\alpha, \mu)\big) = \alpha. \tag{4}$$

$Q(\alpha, \mu)$ is just the α quantile of τ under μ. It follows from (3) and (4) that

$$R\big(Q(\alpha, \mu), \mu\big) = \alpha, \quad \text{and, conversely,} \quad Q\big(R(\alpha, \mu), \mu\big) = \alpha, \tag{5}$$

from which it is clear that, for given μ, R and Q are inverse functions.

The bootstrap test rejects at nominal level α if $\tau < Q(\alpha, \mu^*)$, that is, if τ is smaller than the the α-quantile of the bootstrap DGP. By acting on both sides with $R(\cdot, \mu^*)$, this condition can also be expressed as

$$R(\tau, \mu^*) < R\big(Q(\alpha, \mu^*), \mu^*\big) = \alpha.$$

This makes it clear that the bootstrap P value is just $R(\tau, \mu^*)$. It follows that, if R actually depends on μ^*, that is, if τ is not an exact pivot, the bootstrap test is not equivalent to the asymptotic test, because the former depends not only on τ, but also on the random μ^*.

In Davidson and MacKinnon (1999), it is shown that bootstrap tests enjoy a further refinement, over and above that due to the use of an asymptotic pivot, if τ and μ^* are asymptotically independent. In addition, such asymptotic independence makes it possible to obtain an approximate expression for the size distortion of a bootstrap test. Suppose first that τ and μ^* are fully independent under the true DGP μ_0. Then the rejection probability under μ_0 of the bootstrap test at nominal level α is

$$\begin{aligned}\Pr_{\mu_0}\big(\tau < Q(\alpha, \mu^*)\big) &= E_{\mu_0}\Big(\Pr_{\mu_0}\big(\tau < Q(\alpha, \mu^*) \;\big|\; \mu^*\big)\Big) \\ &= E_{\mu_0}\Big(R\big(Q(\alpha, \mu^*), \mu_0\big)\Big).\end{aligned}$$

Let the random variable q be defined by

$$q = R\big(Q(\alpha, \mu^*), \mu_0\big) - R\big(Q(\alpha, \mu_0), \mu_0\big) = R\big(Q(\alpha, \mu^*), \mu_0\big) - \alpha. \tag{6}$$

This random variable depends on the true DGP μ_0 and the nominal level α. In terms of q, the rejection probability is

$$E_{\mu_0}(q + \alpha) = \alpha + E_{\mu_0}(q). \tag{7}$$

This is an exact result if τ and μ^* are independent. It is useful because, as we see in section 4, $E_{\mu_0}(q)$ can be estimated easily by simulation.

3. The Wild Bootstrap

Consider the linear regression model

$$y_t = x_{t1}\beta_1 + \boldsymbol{X}_{t2}\boldsymbol{\beta}_2 + u_t, \quad t = 1, \ldots, n, \tag{8}$$

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 error terms 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 error terms are assumed to be mutually independent and to have a common mean of zero, but they may be heteroskedastic, with $E(u_t^2) = \sigma_t^2$. 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. With such heteroskedasticity, the usual t statistic is not even asymptotically pivotal for model (8), since its distribution depends on the pattern of the σ_t^2. For both linear and nonlinear regression models, some variant of the Eicker-White heteroskedasticity consistent covariance matrix estimator, or HCCME, can be used

for asymptotically correct inference on the parameters of the regression function in the presence of heteroskedasticity of unknown form.

Both the usual t statistic and asymptotic t statistics based on an HCCME are pivotal if the model is restricted in such a way that the σ_t^2 are fixed. In other words, these statistics are (exactly) pivotal with respect to variations in the values of the regression parameters β_1 and $\boldsymbol{\beta}_2$, but not with respect to the skedastic parameters σ_t^2. On the other hand, HCCME-based statistics are asymptotically pivotal with respect to all the parameters of (8), and all possible error distributions satisfying mild regularity conditions, which must at least include the existence of the variances of these distributions.

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 (8) is

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{9}$$

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. We refer to the version (9) 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 error variance, gives rise to form HC_1 of the HCCME. In the homoskedastic case, the variance of $\hat{u}_t$ is $1 - h_t$, 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 $\boldsymbol{X}$. Normalising by this variance 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 to propose form HC_3, for which the $\hat{u}_t$ are replaced by $\hat{u}_t/(1-h_t)$. MacKinnon and White, and Chesher and Jewitt, show that 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, the problem with bootstrapping a t statistic based on any HCCME is that, since the heteroskedasticity is of unknown form, it cannot be mimicked in the bootstrap distribution. The wild bootstrap gets round this problem as follows. In order to generate a bootstrap sample, we use a bootstrap DGP such that, for $t = 1, \ldots, n$,

$$y_t^* = \boldsymbol{X}_t \hat{\boldsymbol{\beta}} + u_t^*, \tag{10}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of OLS parameter estimates, and the bootstrap error terms are given by

$$u_t^* = f_t(\hat{u}_t) \varepsilon_t, \tag{11}$$

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, with CDF F, defined so as to satisfy

$$E(\varepsilon_t) = 0 \qquad \text{and} \qquad E(\varepsilon_t^2) = 1. \tag{12}$$

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, a further condition is often added to those which F must satisfy, namely that $E(\varepsilon_t^3) = 1$. Liu (1988) considers model (8) with $k = 1$, and shows that 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 $O(n^{-1})$. Mammen (1993) also imposes the extra condition, but his problem is somewhat different from the one dealt with here, in that he uses conventional F statistics rather than HCCME-based statistics, and is not concerned with bootstrap refinements. One of his suggestions for the distribution of the ε_t is probably the most popular choice in recent literature on the wild bootstrap. It is the following two-point distribution:

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

Liu mentions another possibility:

$$F_2: \quad \varepsilon_t = \begin{cases} 1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5, \end{cases} \tag{13}$$

which, for the case she considers, satisfies necessary conditions for refinements in the case of unskewed error terms. Unfortunately, she does not follow up this possibility, since (13), being a lattice distribution, does not lend itself to techniques based on Edgeworth expansion. The techniques used in this paper will allow us to show that (13) 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 (11) by

$$u_t^* = f_t(|\hat{u}_t|)\varepsilon_t, \tag{14}$$

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 (10) clearly belongs to the null hypothesis if the first component of $\hat{\boldsymbol{\beta}}$, corresponding to the regressor $\boldsymbol{x}_1$, is set equal to zero. The regression function has the correct form, and the error terms have mean zero, and are heteroskedastic, for both formulations, (11) or (14), and with any distribution for the ε_t satisfying (12). Since any of the HCCME-based statistics we have discussed is asymptotically pivotal, inference based on the wild bootstrap using such a statistic applied to model (8) should be asymptotically valid. In fact, in the linear case, it is possible to simplify (10) further, since, except for the exogenous regressors, the test statistics depend only on the OLS residuals, and so their distributions are independent of $\boldsymbol{\beta}_2$. Similarly, there is no loss of generality in testing a zero restriction on β_1 rather than testing for some nonzero value. In the case of a nonlinear regression, of course, the distribution of the test statistic depends 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. For (10), that is the case for $\hat{\boldsymbol{\beta}}_2$, but not for the $\hat{u}_t$. Although the HCCME is consistent for the covariance matrix of the OLS estimator, its diagonal elements, based on the $\hat{u}_t^2$, are not consistent estimators of the σ_t^2.

On the other hand, the additional refinement discussed by Davidson and MacKinnon (1999), based on the asymptotic independence of the statistic τ and the bootstrap DGP μ^*, is available for the wild bootstrap, if some precautions are taken. As discussed in that paper, an essential step in achieving this asymptotic independence is to base μ^* exclusively on estimates *under the null hypothesis*. Thus (10) becomes just

$$y_t^* = u_t^*, \qquad u_t^* = f_t(\tilde{u}_t)\varepsilon_t, \tag{15}$$

where the OLS residuals $\tilde{u}_t$ are obtained from the regression

$$y_t = \boldsymbol{X}_{t2}\boldsymbol{\beta}_2 + u_t$$

that corresponds to the null hypothesis. The transformation f may involve taking the absolute value of the argument. It is not only convenient but desirable to use the restricted residuals $\tilde{u}_t$ not only for the bootstrap DGP, but also in the construction of the HCCME.

We now show that, if the model (8) is restricted so that the error terms have distributions that are symmetric about the mean of zero, all HCCME-based t statistics for $\beta_1 = 0$ are asymptotically independent of some versions of the wild bootstrap DGP (15). First, an easy Lemma.

> **Lemma 1:** A mean-zero random variable u which has zero probability mass on the origin and the density of which is symmetric about the origin is the product of two independent random variables: the absolute value $|u|$ and the sign $\text{sgn}(u)$.

Proof: Denote the density of u by $f(u)$. Since $f(-u) = f(u)$, the density of $|u|$ is $g(|u|) \equiv 2f(|u|)$. The density of $\text{sgn}(u)$ can be written in terms of indicator functions as $0.5I(u < 0) + 0.5I(u > 0) = 0.5$. Here we use the fact that there is no positive probability mass on the origin itself. The product of the two densities is $2f(|u|)\,.\,0.5 = f(|u|) = f(u)$, the density of u itself. The factorisation of this density shows that $|u|$ and $\text{sgn}(u)$ are independent. ∎

> **Theorem 1:** Consider the linear regression model
>
> $$y_t = x_{t1}\beta_1 + \boldsymbol{X}_{t2}\boldsymbol{\beta}_2 + u_t, \tag{16}$$
>
> where the regressors are strictly exogenous, and the error terms are mutually independent with mean zero and distributions symmetric about the origin with no positive probability mass on the origin. The statistic based

on the form HC_0 of the HCCME, computed using restricted residuals, for the hypothesis that $\beta_1 = 0$ can be written as

$$\tau \equiv \boldsymbol{x}_1^\top \boldsymbol{M}_2 \boldsymbol{y} / (\boldsymbol{x}_1^\top \boldsymbol{M}_2 \tilde{\boldsymbol{\Omega}} \boldsymbol{M}_2 \boldsymbol{x}_1)^{1/2}. \tag{17}$$

Here $\boldsymbol{y}$ is the n-vector with typical element y_t, $\tilde{\boldsymbol{\Omega}}$ is the $n \times n$ diagonal matrix with typical diagonal element the square of $\tilde{u}_t$, the t^{th} residual from the OLS estimation of the restricted regression

$$y_t = \boldsymbol{X}_{t2} \boldsymbol{\beta}_2 + u_t,$$

and $\boldsymbol{M}_2 = \mathbf{I} - \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top$ is the orthogonal projection matrix on to the orthogonal complement of the span of the columns of $\boldsymbol{X}_2$.

If the regressors obey the usual regularity condition that $n^{-1} \boldsymbol{X}^\top \boldsymbol{X}$ tends as $n \to \infty$ to a deterministic positive definite finite matrix, and if, in addition, there exist positive bounds $\underline{\sigma}^2$ and $\bar{\sigma}^2$ such that $\underline{\sigma}^2 \leq \sigma_t^2 \leq \bar{\sigma}^2$ for all t, then the statistic τ of (17) is asymptotically independent, under the null hypothesis, of the absolute values $|\tilde{u}_t|$ of the restricted residuals, and consequently also of the wild bootstrap DGP μ^* defined by (15) if the transformation f depends only on the absolute value of its argument.

Proof: By the Frisch-Waugh-Lovell theorem (see, for instance, Davidson and MacKinnon (1993), Chapter 1), the OLS estimate of β_1 from (16) is the same as the OLS estimate from the regression $\boldsymbol{M}_2 \boldsymbol{y} = \boldsymbol{M}_2 \boldsymbol{x}_1 \beta_1 +$ residuals, that is,

$$\hat{\beta}_1 = (\boldsymbol{x}_1^\top \boldsymbol{M}_2 \boldsymbol{x}_1)^{-1} \boldsymbol{x}_1^\top \boldsymbol{M}_2 \boldsymbol{y}. \tag{18}$$

The form HC_0 of the HCCME of the variance of $\hat{\beta}_1$ is obtained by applying (9) to the regression (18). The estimated variance is thus

$$\boldsymbol{x}_1^\top \boldsymbol{M}_2 \tilde{\boldsymbol{\Omega}} \boldsymbol{M}_2 \boldsymbol{x}_1 (\boldsymbol{x}_1^\top \boldsymbol{M}_2 \boldsymbol{x}_1)^{-2}, \tag{19}$$

where the diagonal elements of $\tilde{\boldsymbol{\Omega}}$ are the squares of the residuals from the restricted regression, that is, the squares of the elements of the vector $\boldsymbol{M}_2 \boldsymbol{y}$. Equation (17) follows from (18) and (19).

Because $\tilde{\boldsymbol{\Omega}}$ is diagonal, we can express the matrix product $\boldsymbol{x}_1^\top \boldsymbol{M}_2 \tilde{\boldsymbol{\Omega}} \boldsymbol{M}_2 \boldsymbol{x}_1$ as

$$\sum_{t=1}^{n} (\boldsymbol{M}_2 \boldsymbol{x}_1)_t^2 (\boldsymbol{M}_2 \boldsymbol{y})_t^2.$$

Under the null, $\boldsymbol{M}_2 \boldsymbol{y} = \boldsymbol{M}_2 \boldsymbol{u}$, and so the statistic (17) can be written as

$$\frac{\sum_{t=1}^{n} (\boldsymbol{M}_2 \boldsymbol{x}_1)_t (\boldsymbol{M}_2 \boldsymbol{u})_t}{\left(\sum_{t=1}^{n} (\boldsymbol{M}_2 \boldsymbol{x}_1)_t^2 (\boldsymbol{M}_2 \boldsymbol{u})_t^2 \right)^{1/2}} \tag{20}$$

Let us write $(\boldsymbol{M}_2 \boldsymbol{u})_t = |\tilde{u}_t| s_t$, where s_t, the sign of $\tilde{u}_t$, is equal to either $+1$ or -1. In addition, let us write $z_t \equiv (\boldsymbol{M}_2 \boldsymbol{x}_1)_t |\tilde{u}_t|$. Then (20) becomes

$$\frac{\sum_{t=1}^{n} z_t s_t}{\left(\sum_{t=1}^{n} z_t^2 \right)^{1/2}}. \tag{21}$$

Asymptotically, the residuals $\tilde{u}_t$ are equal to the error terms u_t. Thus asymptotically, the statistic is equivalent to (21) with z_t and s_t defined using the u_t instead of the $\tilde{u}_t$. In that case, the z_t and the s_t are independent, by Lemma 1, and so, conditional on the z_t, the s_t are mutually independent and distributed according to the law F_2 of (13). Under the regularity conditions of the second part of the theorem, the central limit theorem can be applied to show that the asymptotic distribution of (21) conditional on the z_t is standard normal. Since this asymptotic distribution is independent of the z_t and so of the $|\tilde{u}_t|$, it follows that τ is asymptotically independent of the $|\tilde{u}_t|$, and so also of any μ^* defined exclusively in terms of the exogenous regressors and the $|\tilde{u}_t|$. ∎

Remarks and Corollaries: Note that the theorem applies to any wild bootstrap defined by (15) and based on the absolute values of the residuals, whatever may be the distribution of the ε_t satisfying (12). If this distribution is itself symmetric about the origin, like F_2, then the theorem applies with any transformation f that is either even or odd, because then the vector with typical element $f_t(\tilde{u}_t)\varepsilon_t$ has the same distribution as that with typical element $f_t(|\tilde{u}_t|)\varepsilon_t$. We refer to any wild bootstrap DGP that can be expressed, implicitly or explicitly, in terms of the $|\tilde{u}_t|$ only as a *tamed* wild bootstrap DGP.

It is easy to adapt the above proof so that it applies to the case in which a joint hypothesis is tested with more than one degree of freedom. The statistic takes on a chi-squared form. If we replace the single column $\boldsymbol{x}_1$ by a matrix $\boldsymbol{X}_1$, we may define a matrix $\boldsymbol{Z}$ with typical row $(\boldsymbol{M}_2\boldsymbol{X}_1)_t|\tilde{u}_t|$, and the n-vector $\boldsymbol{s}$ with typical element s_t. The statistic becomes

$$\boldsymbol{s}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{s} = \boldsymbol{s}^\top \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{s}, \tag{22}$$

where $\boldsymbol{P}_{\boldsymbol{Z}}$ is the orthogonal projection on to the columns of $\boldsymbol{Z}$. Since $\boldsymbol{s}$ and $\boldsymbol{Z}$ are asymptotically independent, the asymptotic distribution of (22) conditional on the $|\tilde{u}_t|$ is chi-squared with as many degrees of freedom as $\boldsymbol{X}_1$ has columns, and, since this does not depend on the $|\tilde{u}_t|$, the statistic is asymptotically independent of the $|\tilde{u}_t|$ and so of any tamed bootstrap DGP μ^*.

Since the other versions, HC_1 to HC_3, of the HCCME all depend only on the absolute values of the residuals, the theorem applies equally well to statistics computed using them. This is particularly obvious for HC_1, which differs from HC_0 only by a multiplicative factor. For HC_2 and HC_3, it suffices to note that the h_t depend only on the exogenous regressors.

In the case of nonlinear regression, the bootstrap DGP must be constructed using a consistent estimate of $\boldsymbol{\beta}_2$. The NLS estimate $\tilde{\boldsymbol{\beta}}_2$ obtained by estimating the restricted model with $\beta_1 = 0$ is asymptotically independent of the NLS residuals $\tilde{u}_t$ from the same regression. A tamed bootstrap DGP can thus be defined exclusively in terms of $\tilde{\boldsymbol{\beta}}_2$ and the $|\tilde{u}_t|$. The statistic τ is then asymptotically independent of this tamed bootstrap DGP.

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.

Theorem 2: Consider the linear regression model

$$y_t = \boldsymbol{X}_t \boldsymbol{\beta} + u_t \tag{23}$$

where the $n \times k$ matrix $\boldsymbol{X}$ with typical row $\boldsymbol{X}_t$ is independent of all the symmetrically distributed error terms u_t, which satisfy the same regularity conditions as for Theorem 1. Under the null hypothesis that $\boldsymbol{\beta} = \mathbf{0}$, the χ^2 statistic for a test of that null against the alternative represented by (23), based on any of the four HCCMEs considered here, has exactly the same distribution as the same statistic bootstrapped, if the bootstrap DGP is the wild bootstrap (15), 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 (13).

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.

Proof: The OLS estimates from (23) 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 (9), with an appropriate choice of $\hat{\boldsymbol{\Omega}}$. The χ^2 statistic thus takes the form

$$\tau \equiv \boldsymbol{y}^\top \boldsymbol{X} (\boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{24}$$

Under the null, $\boldsymbol{y} = \boldsymbol{u}$, and each component u_t of $\boldsymbol{u}$ can be written as $|u_t| s_t$, where $|u_t|$ and s_t are independent. 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. It follows from Lemma 1 that the entire $n \times k$ matrix $\boldsymbol{Z}$ with typical row $\boldsymbol{Z}_t$ is independent of the vector $\boldsymbol{s}$. The statistic (24) becomes

$$\boldsymbol{s}^\top \boldsymbol{Z} \left(\sum_{t=1}^{n} a_t \boldsymbol{Z}_t^\top \boldsymbol{Z}_t \right)^{-1} \boldsymbol{Z}^\top \boldsymbol{s}, \tag{25}$$

where $a_t = 1$ 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

$$\boldsymbol{\varepsilon}^\top \boldsymbol{Z} \left(\sum_{t=1}^{n} a_t \boldsymbol{Z}_t^\top \boldsymbol{Z}_t \right)^{-1} \boldsymbol{Z}^\top \boldsymbol{\varepsilon}, \tag{26}$$

where $\boldsymbol{\varepsilon}$ denotes the vector containing the ε_t. The matrix $\boldsymbol{Z}$ is exactly the same as in (25), because the exogenous matrix $\boldsymbol{X}$ is reused unchanged by the wild bootstrap, and the wild bootstrap error 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 we saw in the proof of Theorem 1 that, under the null hypothesis, $\boldsymbol{s}$ 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. The statistic τ must take on one of the 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, $i = 1, \ldots, 2^n$, 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, $i = 0, \ldots, 2^n - 1$, 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 p^*. ∎

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 $\hat{p}^*$.

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 (25) 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.

In terms of the analysis in section 2, the result of Theorem 2 can be interpreted as the vanishing of the random variable q of (6) for all values of α equal to the dyadic numbers $i/2^n$. Since τ and τ^* have the same distribution, the functions $R(\cdot, \mu_0)$ and $R(\cdot, \mu^*)$ are the same, as are $Q(\cdot, \mu_0)$ and $Q(\cdot, \mu^*)$.

Given the exact result of the theorem, it is of great interest to see the extent of the size distortion of the wild bootstrap with F_2 when the null hypothesis involves only a subset of the regression parameters. This question will be investigated by simulation in the following sections. At this stage, it is possible to see the reason for which the theorem does not apply in that case. The expressions (24) and (25) for τ and τ^* continue to hold if $\boldsymbol{Z}_t$ is redefined as $|\tilde{u}_t|\boldsymbol{X}_t$. 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 error terms $\boldsymbol{u}$. 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. Since, for each t, $\tilde{u}_t$ tends to u_t as $n \to \infty$, the asymptotic independence of $|\tilde{u}_t|$ and s_t still holds, and so the asymptotic distributions of τ and τ^* still coincide.

4. Estimation of Size Distortion

In this section, we see that we can very accurately estimate the difference between the distribution of an HCCME-based statistic τ and that of its wild-bootstrap counterpart τ^* by simulation, if the wild bootstrap is tamed. The key to this is the asymptotic independence of τ and a tamed bootstrap DGP μ^*, given by Theorem 1. Full independence between τ and μ^* allows us to assert that the true rejection probability of a bootstrap test at nominal level α is given by (7), so that the error in rejection probability (ERP) is the expectation, under the true DGP μ_0, of the random variable q given by (6).

The expectation of q can be written as

$$E_{\mu_0}\Big(R\big(Q(\alpha, \mu^*), \mu_0\big)\Big) - \alpha. \tag{27}$$

Now, by (5), $R\big(Q(\alpha, \mu_0), \mu_0\big) = \alpha$. Thus, for given μ_0, (27), considered as a function of α, is a bias function. In the spirit of linear bias correction, we approximate $R(Q(\alpha, \mu^*), \mu_0)$ as an affine function of its first, random, argument, and get

$$R\big(Q(\alpha, \mu^*), \mu_0\big) \approx \alpha + R_1.\big(Q(\alpha, \mu^*) - Q(\alpha, \mu_0)\big),$$

where R_1 is the derivative of R with respect to its first argument evaluated at $(Q(\alpha, \mu_0), \mu_0)$. Similarly,

$$R\big(Q(\alpha, \mu_0), \mu^*\big) \approx \alpha + R_1.\big(Q(\alpha, \mu_0) - Q(\alpha, \mu^*)\big),$$

and so, approximately, the expectation of q is given by

$$\alpha - E_{\mu_0}\Big(R\big(Q(\alpha, \mu_0), \mu^*\big)\Big). \tag{28}$$

Consider a random variable τ^* of which a drawing under μ_0 is generated as follows. A sample is drawn from μ_0 and used to compute a drawing $\hat{\mu}$ of the bootstrap DGP μ^*. Then a sample is drawn from $\hat{\mu}$, and used to compute a bootstrap statistic, which is then the drawing of τ^*. The CDF of τ^*, evaluated at argument α, can be seen to be just $E_{\mu_0}\big(R(\alpha, \mu^*)\big)$: Conditional on $\hat{\mu}$, the probability that $\tau^* < \alpha$ is given by the CDF of the bootstrap statistic under $\hat{\mu}$, that is, $R(\alpha, \hat{\mu})$. The unconditional expectation of this probability is just $E_{\mu_0}\big(R(\alpha, \mu^*)\big)$, as required.

The above construction allows us to evaluate the expectation in (28) easily by simulation. For each replication, the DGP μ_0 is used to draw realisations of the statistic τ and the bootstrap DGP μ^*. Next the realisation $\hat{\mu}$ of μ^* is used to draw a realisation of τ^*. The quantile $Q(\alpha, \mu_0)$ is then estimated as usual as the α-quantile of the drawings of τ, and the expectation of $R\big(Q(\alpha, \mu_0), \mu^*\big)$ as the proportion of the drawings of τ^* that are less than the estimate of $Q(\alpha, \mu_0)$. This method of estimating the expectation of q was first suggested by Davidson and MacKinnon (1999).

In fact, by a manipulation frequently used in bias correction, a more accurate estimate of the expectation of q can be obtained by a slight modification of the above procedure, in which the roles of the distributions of τ and τ^* are interchanged. Drawings of τ and τ^* are made exactly as described above, but then the expectation of q is estimated as the proportion of drawings of τ less than the α-quantile of τ^*, minus α. We show in the next section that this procedure does indeed yield better estimates of the size distortion of the bootstrap test; good enough to be almost indistinguishable, up to experimental error, from brute force estimates obtained by experiments in which a complete bootstrap test is undertaken for each replication.

It is clear that, in practice, it is not necessary to convert test statistics to approximate P value form in order to estimate size distortions by the above procedure. Drawings of the statistics are obtained in whatever form is most convenient, and sorted in order from the most extreme values to the least extreme. For each value of α of interest, it is then straightforward to compute the proportion of realisations of the statistic more extreme than the realisation of the bootstrap statistic in position α in the sorted list.

5. Experimental Design and Simulation Results

In this section, we use the procedure outlined in the previous section to study the extent of the size distortion of wild bootstrap tests based on statistics computed with an HCCME. First, we check that we can indeed obtain results from this procedure that are almost experimentally indistinguishable from those we obtain by much more laborious, but theoretically irreproachable, simulation methods. Having established that, we move on to investigate how the size distortion depends on all those aspects of the problem on which it might depend, with a view to estimating the greatest distortion to which a wild bootstrap test might be subjected in various circumstances. A point of particular importance is to compare the performance of wild bootstrap tests that use the asymmetric F_1 distribution and the symmetric F_2 distribution.

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. This implies that HCCME-based test statistics will be farthest from being pivotal, that is, most dependent on the unknown aspects of the true DGP, when there is severe heteroskedasticity and when there are observations with high leverage. Further,

one expects bootstrap tests to behave better in large samples than in small. Thus, in order to stress-test the wild bootstrap, and along with it the procedure of the previous section, most of our experiments are performed with a sample of size 10, which contains one regressor all the elements but one of which are independent drawings from $N(0,1)$, but the second of which is 10, so as to create an observation with exceedingly high leverage. In Table 1, the components of this regressor, which we denote by $\boldsymbol{x}_1$, are given, along with those of four other regressors, $\boldsymbol{x}_i$, $i = 2, \ldots, 5$, used in the experiments. 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 the constant, and $\boldsymbol{x}_1$, the constant, and increasing numbers of other regressors $\boldsymbol{x}_i$, $i = 2, \ldots, 5$. The h_t measure the leverage of the 10 observations for the different regression designs.

The results of two experiments are shown in Figure 2. For both, samples are drawn from the following DGP, which plays the role of μ_0 in the experiments:

$$y_t = \sigma_t u_t, \qquad t = 1, \ldots, 10, \tag{29}$$

where $\sigma_t^2 = x_{t1}^2$, the t^{th} component of $\boldsymbol{x}_1$, and the u_t are normal white noise. This pattern of heteroskedasticity leads to bias of the OLS covariance matrix; see White (1980). Since $\boldsymbol{x}_1$ contains a very high leverage observation for $t = 2$, the DGP (29) is very strongly heteroskedastic. Note that, since the distributions of all the statistics we consider are independent of the parameters $\boldsymbol{\beta}$ of the regression, we may, as in (29), set $\boldsymbol{\beta} = \mathbf{0}$ without loss of generality.

The statistic bootstrapped, τ, is an HCCME-based pseudo-t statistic on $\boldsymbol{x}_1$ in the model

$$\boldsymbol{y} = \beta_0 \boldsymbol{\iota} + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \boldsymbol{u},$$

where $\boldsymbol{\iota}$ is a vector of ones representing the constant. The null hypothesis under test can be written as

$$\boldsymbol{y} = \beta_0 \boldsymbol{\iota} + \beta_2 \boldsymbol{x}_2 + \boldsymbol{u}. \tag{30}$$

(This is the case labelled $k = 3$ in Table 2.) The residuals $\tilde{u}_t$ from (30) are used to compute version HC_2 of the statistic (17), in which in this case $\boldsymbol{M}_2$ projects off the constant $\boldsymbol{\iota}$ and $\boldsymbol{x}_2$. The h_t used in the statistic are the diagonal elements of the complementary projection $\boldsymbol{P}_2$. The bootstrap DGP μ^* is then a wild bootstrap with distribution F_2 (which tames it):

$$y_t^* = \tilde{u}_t (1 - h_t)^{-1/2} \varepsilon_t, \quad \varepsilon_t \sim F_2. \tag{31}$$

For the first experiment, $N = 100,000$ replications were performed on each of which drawings were made of τ and μ^*, and then a drawing of the variable τ^*, as described in the previous section. For each $\alpha = 0.01, 0.02, \ldots, 0.99$, the expectation of the ERP of the bootstrap test was calculated approximately by the two procedures of the previous section, in which the empirical distributions of the drawings of τ and τ^* are used. For the second experiment, an actual bootstrap P value was computed on each replication, with 399 bootstrap samples.

The results of the two experiments are shown in Figure 2 using P value discrepancy plots, as described in Davidson and MacKinnon (1998). These plots show, as

a function of the nominal level α, the difference between the true rejection probability and the nominal level. Since the statistic has just one degree of freedom, 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 it is known (see, for instance. Hall (1992)) that, in many circumstances, the errors in the rejection probabilities 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 for the one-tailed test: the rejection region becomes those P values that are too close either to zero or to unity. For the experiments of Figure 2, it is easy to see that, since the u_t are symmetrically distributed, the test statistic is also symmetrically distributed, and so twice the ERP of a one-tailed test at nominal level α is equal to the ERP of a two-tailed test at nominal level 2α. It can be seen that, except possibly for a significant difference for very small values of α, the curves generated by the brute-force procedure and by the second, rapid, procedure based on τ and τ^* are very close. On the other hand, the first of the rapid procedures can be seen to give significantly different results, although the overall shape of the curve is the same as for the other two.

The case treated in Figure 2 is representative of moderate size distortion. We conducted other experiments with both very small and very large ERPs. In all cases, the results were similar to those in Figure 2. For the remainder of our study, therefore, we show results generated by the second rapid method whenever it is possible to do so, that is, whenever the test statistic and the bootstrap DGP are asymptotically independent.

Some preliminary results are shown in Figure 3, where P value discrepancy plots are given for the conventional t statistic, based on the OLS covariance matrix estimate, the four versions of HCCME-based statistics, HC_i, $i = 0, 1, 2, 3$, and a wild bootstrap test based on the HC_3 form of the statistic. P values for the asymptotic tests are obtained using Student's t distribution with 7 degrees of freedom. To avoid redundancy, the plots are drawn only for the range $0 \leq \alpha \leq 0.5$, since the statistic is symmetrically distributed. It can be seen that the HC_i statistics have too little mass in the tails, too much in the midrange, and too little in the centre of the distribution, giving rise to the oscillations clearly visible in the plots. The bootstrap test has different, considerably better, behaviour. The conventional t statistic, which does not have even an asymptotic justification, is of course the worst behaved of all, with far too much mass in the tails.

We now set out to try to answer the following series of questions:

(1) If we use the bootstrap, does it matter which of the four versions of the HCCME is used? It is easy to see that it does not for $k = 1, 2$, but, for $k = 3, 4, 5, 6$, the HC_3 version is markedly better than the other three.

(2) 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, as is the case for the results in Figure 2. The latter, at least for HC_3, seems to give slightly better results in most cases, but not all.

(3) Is it the case, as we might think from (22), that in order to compute the bootstrap statistic, it is best to reuse the $(\boldsymbol{M}_2\boldsymbol{x}_1)_t|\tilde{u}_t|$ unchanged, without generating residuals $\boldsymbol{M}_2\boldsymbol{y}^*$ for the computation of the bootstrap HCCME? (These residuals were in fact generated for the experiments of Figure 2.) It turns out that, unless $k = 1$ in which case it does not matter, this is *not* a good idea.

(4) If the null hypothesis is that $\beta_i = 0$ for $i \neq 0, 1$, rather than that $\beta_1 = 0$, what impact has this on the bootstrap test's performance? As one might expect, the size distortion is usually worse, but not always, for the test of $\beta_1 = 0$ than for the other tests.

(5) All the theory of this paper applies to the case of symmetrically distributed error terms. How is performance affected if the error terms are in fact highly skewed? We investigate the case of errors that follow a centred chi-squared distribution with 2 degrees of freedom. In some cases, performance deteriorates very substantially, and size distortion can remain enough to preclude reliable inference even for $n = 100$.

(6) What is the penalty for using the wild bootstrap when the errors are homoskedastic and inference based on the conventional t statistic is reliable, at least with normal errors? Fortunately, the answer is that the penalty is very small.

(7) How is performance affected if the leverage of observation 2 is reduced? We expect the underlying statistic to be closer to pivotal when there is less leverage, and indeed the ERP of the bootstrap test is often substantially less in this case, but not uniformly so. The presence or absence of heteroskedasticity seems to be a much more important determinant of the ERP than the presence or absence of leverage.

(8) How quickly does the ERP of the bootstrap test tend to zero as the sample size n grows? The usual theory based on Edgeworth expansions does not give a clear answer to this question, since, as we saw in section 3, Beran's standard argument does not apply, since we cannot estimate the σ_t^2 consistently. Our experiments show that, with extreme heteroskedasticity, inference becomes reliable only slowly as n grows. Even for $n = 100$, there remains significant size distortion.

(9) The wild bootstraps that we have considered based on distribution F_2 are automatically tamed, whether or not the error terms are symmetrically distributed, because all the transformations f_t used by the HC_i, $i = 0, \ldots, 3$, are odd. That based on F_1 is not unless the transformation f_t depends only on the absolute value of its argument. Although signed residuals are usually used for wild bootstraps based on F_1, it is perfectly possible to replace them by their absolute values, thereby taming them. Is it then the case that, as theory suggests, the ERP of the tamed bootstrap tends to zero more quickly as the sample size grows? The answer is yes, but, once more, there remains substantial size distortion even for $n = 100$, and more so than with the F_2-based bootstrap.

Finally, we wish to answer all these questions for the asymmetric wild bootstrap based on F_1. It turns out that, in most cases, the two wild bootstraps, based on

F_1 and F_2, have not too dissimilar behaviour. In all cases we consider, that based on F_2 has the smaller ERP, although sometimes not significantly so. In particular, even when the error terms are highly skewed, use of F_1 does not seem to yield any advantage.

6. Experimental Results

The basic results that serve as a benchmark for all others are shown in Figure 4. P value discrepancy plots are drawn for the wild bootstrap test of $\beta_1 = 0$, with the symmetric F_2 distribution and the HC_3 transformation $f_t(u_t) = u_t/(1-h_t)$, based on the HC_3 version of the HCCME, where the bootstrap version of the statistic uses residuals, $\boldsymbol{M}_2\boldsymbol{y}^*$ rather than the unchanged $(\boldsymbol{M}_2\boldsymbol{x}_1)_t|\tilde{u}_t|$. The regressors are those of Table 1, and the dependent variable is generated by the DGP (29), that is, with heteroskedastic normal error terms. All experiments used 100,000 replications and generated bootstrap ERPs by the second rapid procedure of section 4. It can be seen that the ERP of the bootstrap test is very sensitive to specific features of the regression design. As expected, the ERP for $k = 1$ is just experimental noise, but, for higher values of k, the ERP is significantly different from zero, but not with any clear pattern. It is small for $k = 2$, and becomes substantially greater for $k = 3$ and especially for $k = 4$. By what is presumably a coincidence induced by the specific form of the data, the ERP for $k = 5$ is much smaller, and that for $k = 6$ is barely larger. Although the simulation results are quite clear, it is not obvious just how to characterise analytically the determinants of the ERP.

Question (1) of the previous section is treated in Figure 5. Panels a and b are completely analogous to Figure 4, except that HCCMEs of type HC_0 and HC_2 are used instead of HC_3. It is unnecessary to draw plots for HC_1 for bootstrap purposes, since the HC_0 and HC_1 statistics differ only by a constant multiplicative factor, and so the bootstrap P values are identical for the two statistics. In order to facilitate comparison of the three distinct versions of the statistic, panel c shows the plots for all three for $k = 4$, which is the worst behaved case. For both $k = 1$ and $k = 2$, all versions yield identical results. For $k = 1$ this is obvious, since the raw statistics are identical. For $k = 2$, the only regressor other than $\boldsymbol{x}_1$ is the constant, and so h_t does not depend on t. The raw statistics are different, but differ only by a constant multiplicative factor. For $k > 2$, significant differences are clearly visible, and in all cases HC_3 has least distortion.

Some answers to question (2) can be seen in Figure 6. The upper panel (6a) is the analogue of Figure 4, and is obtained with the same setup as that used for that figure, except that the bootstrap DGP is just

$$y_t^* = \tilde{u}_t \varepsilon_t, \quad \varepsilon_t \sim F_2. \tag{32}$$

Implicitly, we set $f_t(u_t) = u_t$. For $k = 1$, there is once more no difference from Figure 4. For other values of k, the differences are not very great, except for $k = 4$. The lower panels, (6b) and (6c), display these differences for the cases $k = 3$ and $k = 4$ respectively. Whereas for $k = 3$ the ERP is slightly smaller with the bootstrap DGP (32), for $k = 4$ we see that there is substantially more

distortion here than in the base case. This would tend to suggest that the base case setup is better, but it is not clear how general such a conclusion would be.

Answers to question (3) are found in Figure 7. The upper panel, (7a), is again the analogue of Figure 4, but the bootstrap procedure is changed so that in the computation of the bootstrap statistics, the $(\boldsymbol{M}_2\boldsymbol{x}_1)_t|\tilde{u}_t|$ are used unchanged, without a further projection of the bootstrap dependent variable $\boldsymbol{y}^*$ by $\boldsymbol{M}_2$ in order to obtain bootstrap residuals $\tilde{u}_t^*$. For $k = 1$ there is of course no difference, since $\boldsymbol{M}_2$ is just the identity matrix, but for other values of k, it is clear that using unprojected residuals leads to substantially different, usually worse, behaviour. The lower panel, (6b), shows the comparison for $k = 3$, a more striking comparison than for $k = 4$, since $k = 4$ is badly behaved even in the base case. Since the norm of the unprojected vector $\tilde{\boldsymbol{u}}$ is greater than that of $\boldsymbol{M}_2\boldsymbol{y}^*$, on average we expect the variance estimate to be greater with the former than with the latter. However, what is true on average is clearly not true in detail. The bootstrap tests with no projection of the residuals underreject severely in the tails of the distribution, but overreject equally severely in the mid-range. There seems no doubt that projecting the residuals is a necessary step in the implementation of the bootstrap test.

Figure 8 addresses question (4). Panel (8a) is as usual the analogue of Figure 4, where the test is of the hypothesis that $\beta_2 = 0$. Results are shown only for $k > 2$, because β_2 does not exist for $k = 0, 1$. It is clear that the distortion is usually less than for the test of $\beta_1 = 0$, (note the scale of the vertical axis) but, for $k = 5$, where that test seems to work almost perfectly, panel (8b) shows that the test for $\beta_2 = 0$ is significantly distorted. Finally, in panel (8c), a comparison is made for $k = 4$ of the test for $\beta_2 = 0$ with HC_0, HC_2 and HC_3. Once again, HC_3 is less distorted than the other two.

In Figure 9, attention is once more focussed on the test for $\beta_1 = 0$. In order to respond to question (5), the error terms were generated using the chi-squared distribution with two degrees of freedom, with the mean of 2 subtracted. This distribution is very highly skewed to the right. Once more, panel (9a) shows the P value discrepancy plots for the HC_3-based bootstrap test, but for the full range of the nominal level α, because with skewed errors, the symmetry about the origin is destroyed, as is clear from the figure. Experimentation showed that the rapid simulation procedure was less reliable than usual with the heavily skewed errors, and so it seemed desirable for this figure to resort to more costly experiments in which a full bootstrap test, with 399 bootstraps, is performed on each of the 100,000 replications. It is clear that the bootstrap test always significantly underrejects. If the errors were skewed to the left, or alternatively if a one-tailed test were performed with rejection to the left rather than to the right, we can see that the tests would overreject. Panel (9b) compares the normal-error case and the skewed-error case for $k = 4$, where although even with normal errors there is substantial distortion, there is, as expected, still more when the errors are skewed. In panel (9c), plots are drawn, still for $k = 4$, with HC_0 and HC_2. With these, it seems that overrejection occurs in the region of interest for one-tailed tests in either direction, although less so in the right-hand tail where HC_3 underrejects.

The case of normal, homoskedastic, errors, mentioned in question (6), is investigated in Figure 10. Since all the statistics considered are scale invariant, we

can without loss of generality set the error variance to 1. In panel (10a), results are shown for all values of k. One can see directly from this panel, and from panel (10b), in which results are compared for homoskedastic and heteroskedastic errors with $k = 4$, that there is little distortion when the errors are homoskedastic. In panel (10c), the three versions of HCCME are compared, again for $k = 4$, and here we see that for once the best performer is HC_2, although even HC_0 is not very seriously distorted. It was possible to use the rapid method for the simulations of Figure 10.

In order to address question (7), results are shown in Figure 11 for the case in which the observation with high leverage is replaced so as to leave an almost balanced design. In panel (11a), the error terms have the same pattern of heteroskedasticity as usual, and in panel (11b) the errors are homoskedastic. Errors are normal in both cases. Panel (11c) compares, for $k = 3$, the four cases with or without heteroskedasticity and with or without leverage. Although the absence of a high-leverage observation reduces the ERP for some values of k, this panel shows clearly that the main cause of size distortion is strong heteroskedasticity. In panel (11d), the same comparison is made for HC_0, for which the same conclusion can be drawn, even more strongly.

Figure 12, designed to respond to question (8), is different from the preceding figures. Here we focus on the worst case observed so far, where the errors are skewed and heteroskedastic, there is a high leverage point, and we test $\beta_1 = 0$. In order to avoid needless distortion, we use version HC_3 of the HCCME. Panel (12a) shows P value discrepancy plots for increasing values of n from 10 up to 100. As n grows, it appears that the distortion lessens, but very slowly. Even for $n = 100$, it remains highly significant, and of a comparable order of magnitude as for $n = 10$. It should be noted that the value of x_{21} is increased along with n in such a way that h_2 remains roughly constant at 0.93, as in the sample for $n = 10$. If x_{21} were kept constant as the sample size grew, then the sample design would become much more balanced for larger n. On the other hand, the value of σ_2 is kept the same as for $n = 10$, whatever the sample size: It does not increase in proportion to x_{21}. In panel (12b), results are shown, just for $n = 10$ and $n = 100$, for the case with normal, but still heteroskedastic error terms. The size distortion is much less, and falls off more rapidly for larger n.

In Figure 13, some answers are provided for question (9). In panel (13a), results are shown for $k = 3$, and in panel (13b) for $k = 4$. P value discrepancies are plotted for HC_3 and the wild bootstrap based on F_1, and the four possible cases in which signed residuals (wild) or absolute values (tame) are used, and errors are symmetric or skewed. With untamed bootstraps, it was necessary to use experiments with full bootstrap tests on each replication. For both symmetric and skewed errors, the two bootstraps have different behaviour, more so when the errors are skewed, but it is hard to say that one is better than the other. Panel (13c) shows the same set of results for $k = 4$ and $n = 100$. While it is clear that the ERP of the tamed bootstrap is less than that of the wild bootstrap, as expected, perhaps the main point to emerge clearly from these figures is that, even for $n = 100$, the distortion is very significant with the F_1-based bootstrap. For the case of symmetric errors, the distortion is in fact greater for $n = 100$ than for $n = 10$.

The results in Figure 13, particularly panel (13b), indicate that the F_1-based bootstrap, wild or tamed, never does as well as the F_2-based bootstrap, which is always tamed by construction. In order to see this in a little more detail, Figure (14) provides in panel (14a) results like those for the base case in Figure 4, but for the tamed F_1-based bootstrap. It can be seen that the ERP is very substantial for $k = 1$ and $k = 2$, where the F_2-based bootstrap gives essentially perfect inference for $k = 1$ and very good inference for $k = 2$. In panel (14b), a comparison is made of the two bootstraps for $k = 3$ and $k = 4$, the two cases in which the behaviour of the F_2-based bootstrap is worst. In these cases also, the F_2-based bootstrap has substantially less size distortion. Panels (14c) and (14d) repeat the exercise with skewed errors, with qualitatively similar results. In particular, the F_2-based bootstrap maintains its better performance, even though distortion remains substantial. It seems reasonable to conclude that the F_2-based bootstrap is never any worse than other variants in all circumstances in which the wild bootstrap is appropriate.

7. Conclusion

The wild bootstrap is commonly applied to models with heteroskedastic error terms and an unknown pattern of heteroskedasticity. Standard results on bootstrap refinements obtainable when the statistic bootstrapped is asymptotically pivotal do not apply to the wild bootstrap, since it is impossible to estimate the unknown pattern of heteroskedasticity consistently. However, another refinement is shown to be available if the statistic that is bootstrapped is asymptotically independent of the bootstrap DGP. If the hypothesis under test is that all the model parameters are zero, we show that, with symmetrically distributed error terms, a version of the wild bootstrap gives exact inference, up to small errors due to a discrete distribution of the bootstrap statistic. The properties of the same version of the wild bootstrap are investigated in various circumstances in which inference is not exact, and it is found that, provided the property of asymptotic independence is satisfied, this version of the wild bootstrap is never any worse behaved than any other, even in cases in which the contrary might be expected on the basis of conventional bootstrap theory. Although our experiments cover a good number of cases, some caution is still necessary on account of the fact that the extent of the size distortion of wild bootstrap tests appears to be very 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 Figures 11 and 12b in particular make 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 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.

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

Davidson, R. and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*, Oxford University Press, New York.

Davidson, R. and J. G. MacKinnon (1999). "The Size Distortion of Bootstrap Tests," *Econometric Theory*, 15, 361–376.

Davidson, R. and J. G. MacKinnon (1998). "Graphical Methods for Investigating the Size and Power of Hypothesis Tests," *The Manchester School*, **66**, 1–26.

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.

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

Horowitz, J. L. (1997). "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in David M. Kreps and Kenneth F. Wallis, eds., *Advances in Economics and Econometrics: Theory and Applications*, Vol. 3, pp. 188–222, Cambridge, Cambridge University Press.

Liu, R. Y. (1988). "Bootstrap procedure 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.

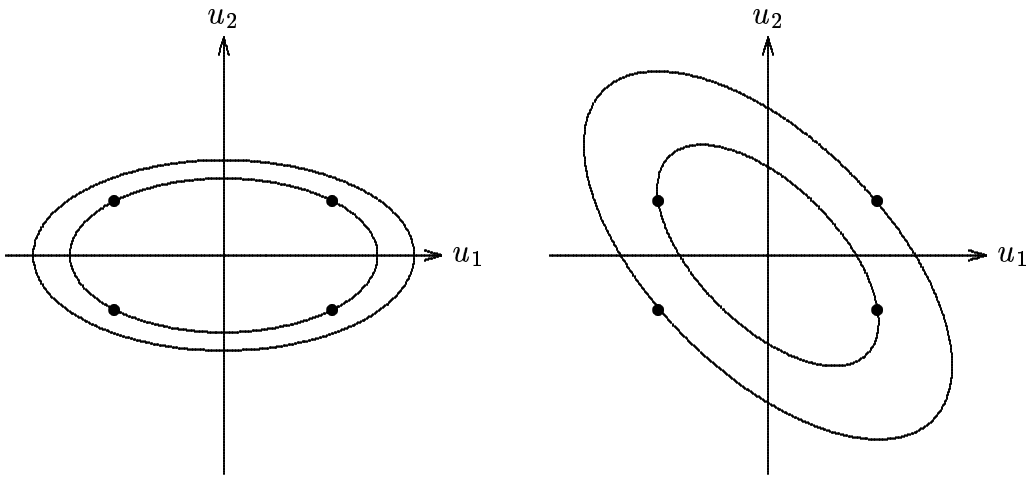

Figure 1. Absolute values and signs of two random variables

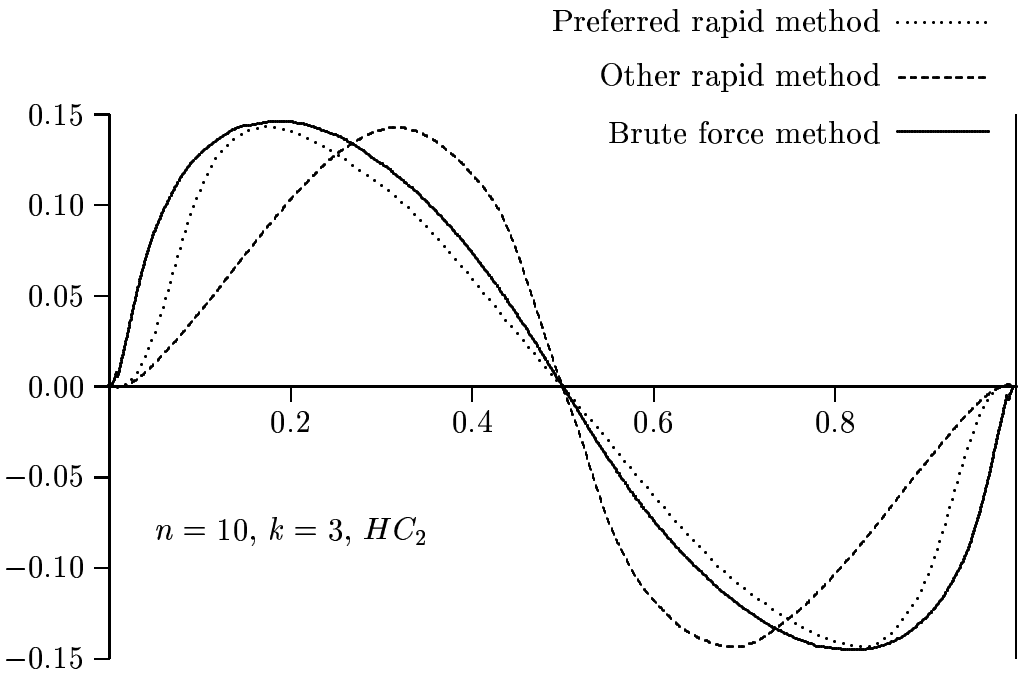

Figure 2. Comparison of brute force and two rapid methods

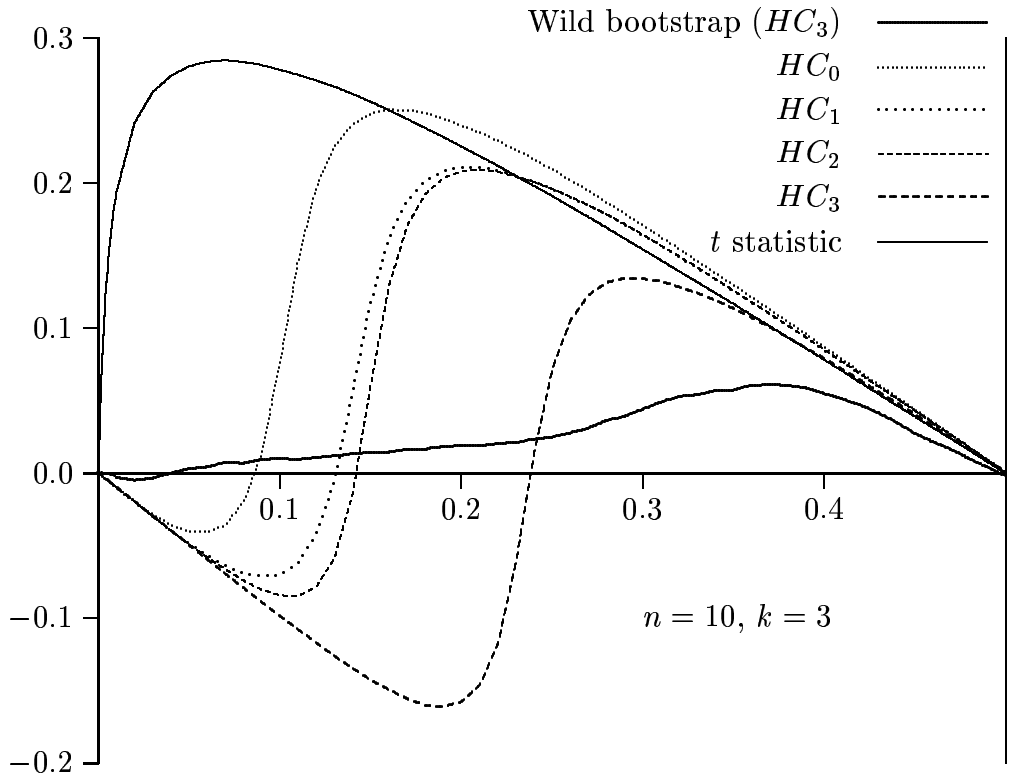

Figure 3. Asymptotic and bootstrap tests

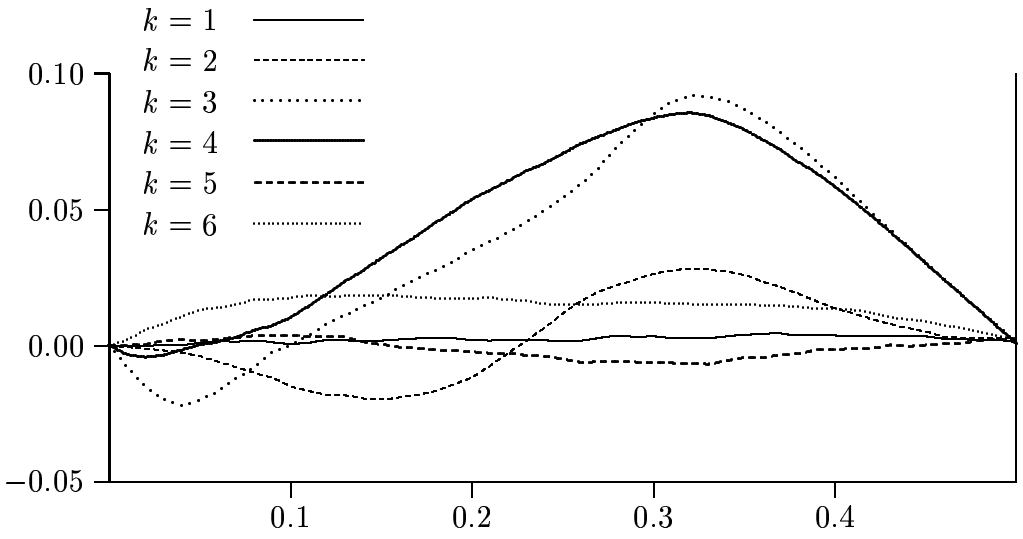

Figure 4. Base case

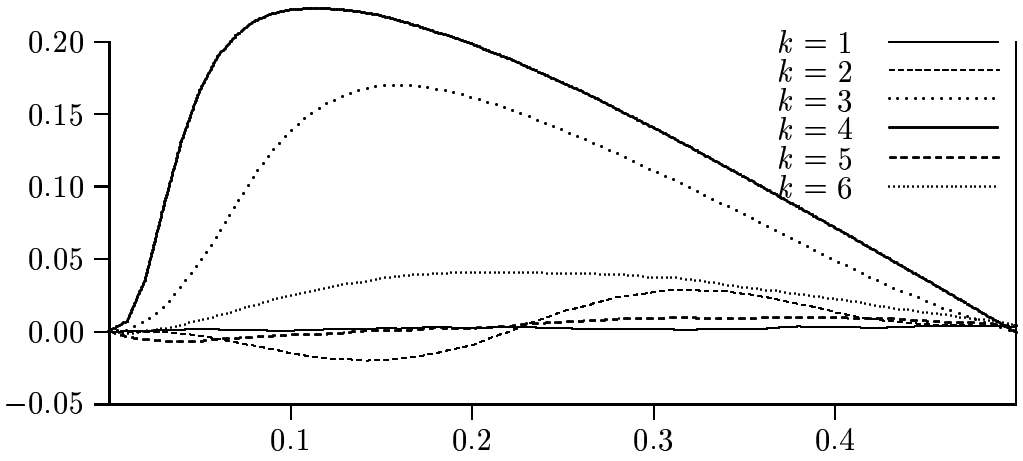

Figure 5a. Base case, but with HC_0

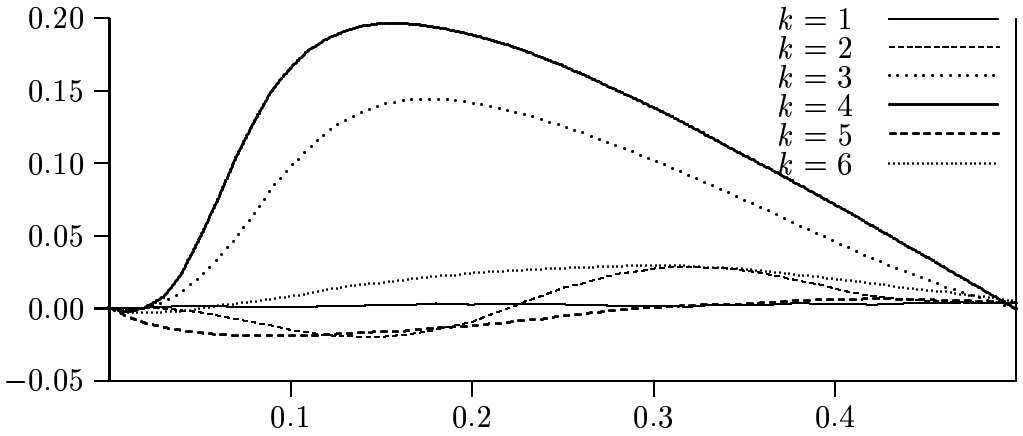

Figure 5b. Base case, but with HC_2

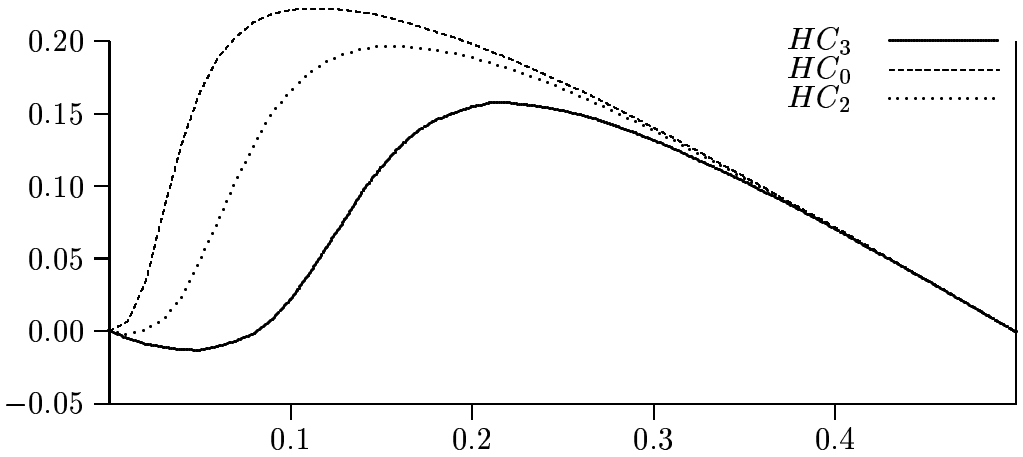

Figure 5c. HC_3 compared with HC_0 and HC_2, $k = 4$

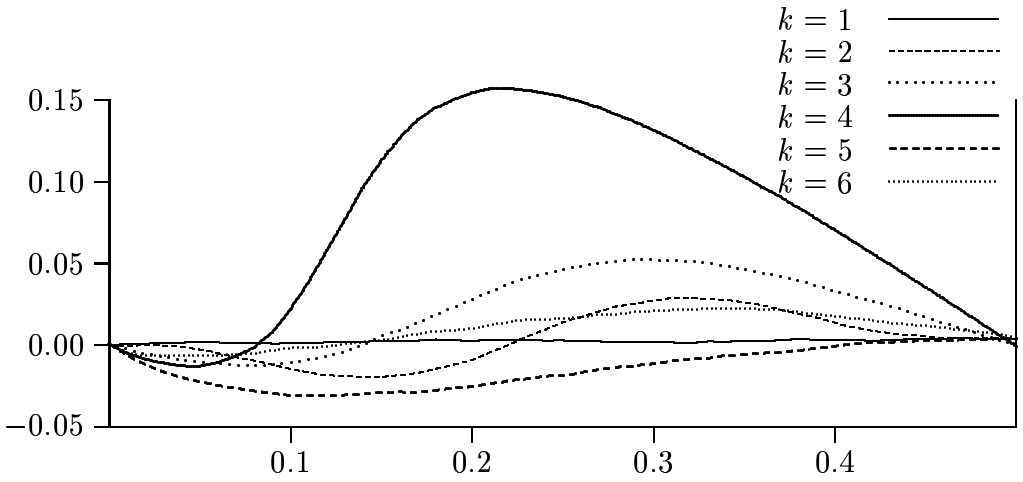

Figure 6a. Base case, but with untransformed residuals

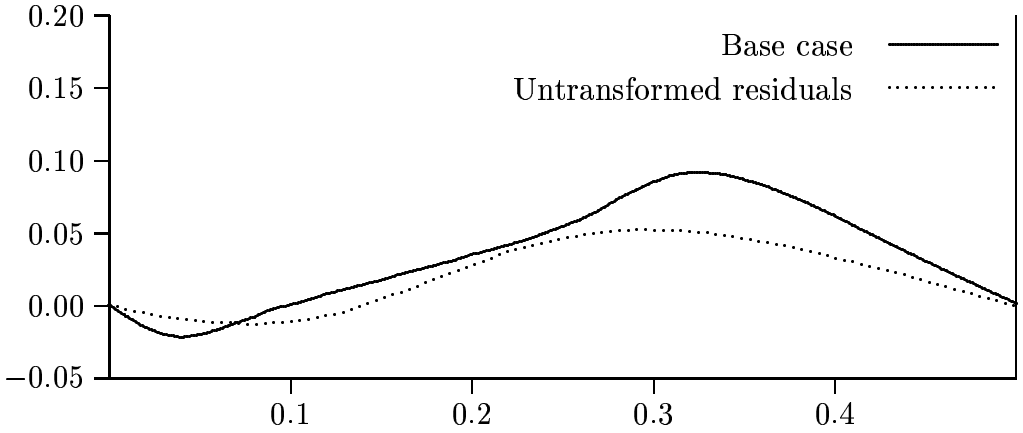

Figure 6b. Comparison with base case, $k = 3$

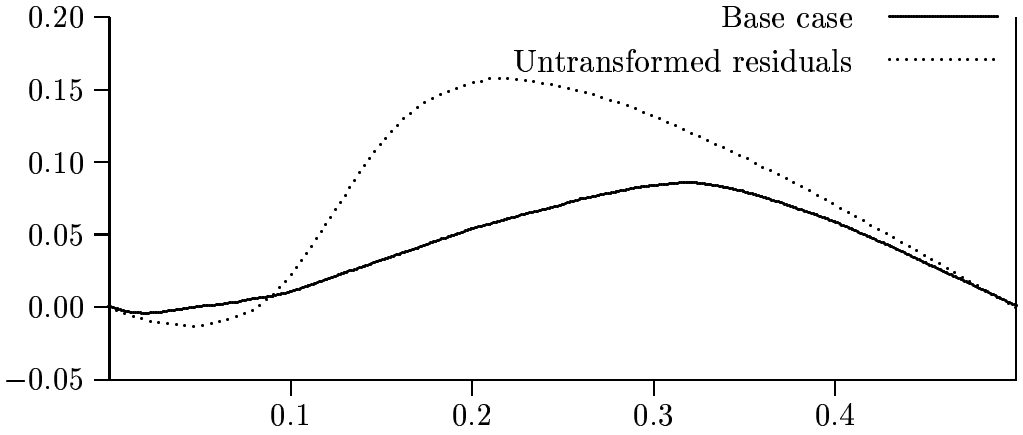

Figure 6c. Comparison with base case, $k = 4$

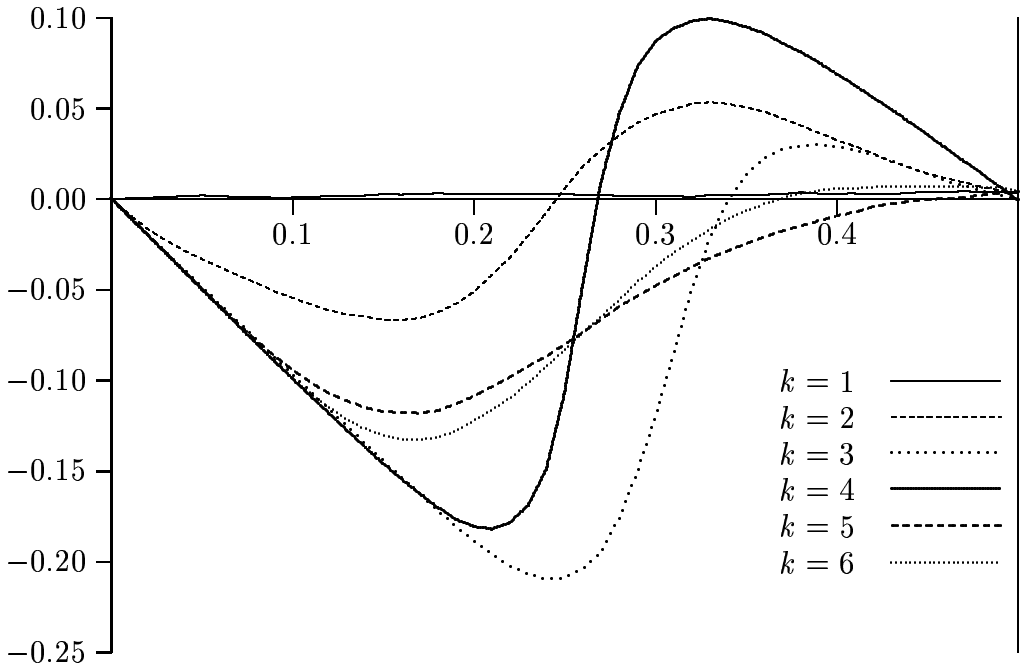

Figure 7a. Base case, but with the $(M_2 x_1)_t |\tilde{u}_t|$ used unchanged

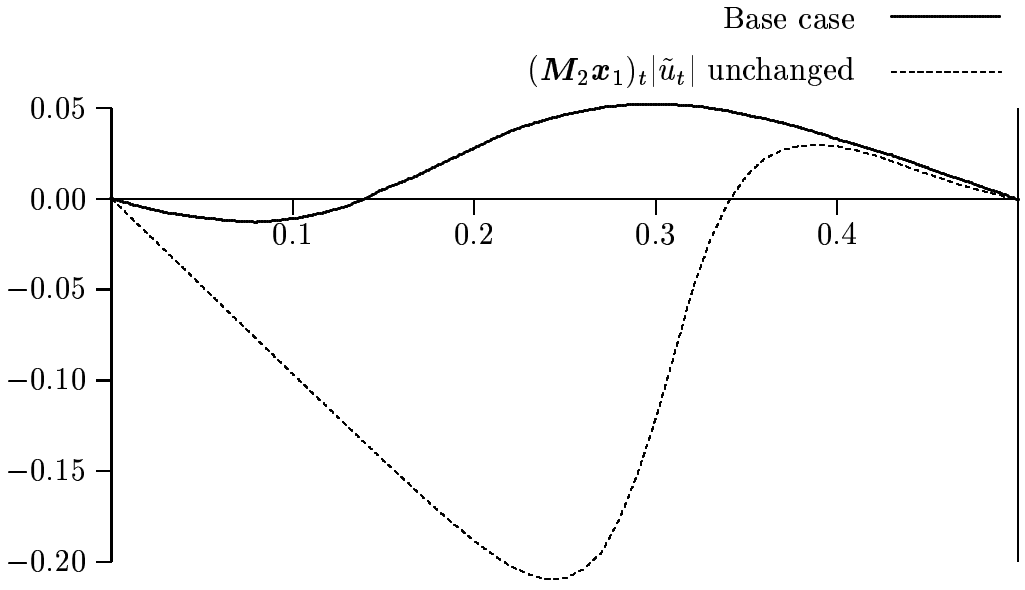

Figure 7b. Comparison with base case for $k = 3$

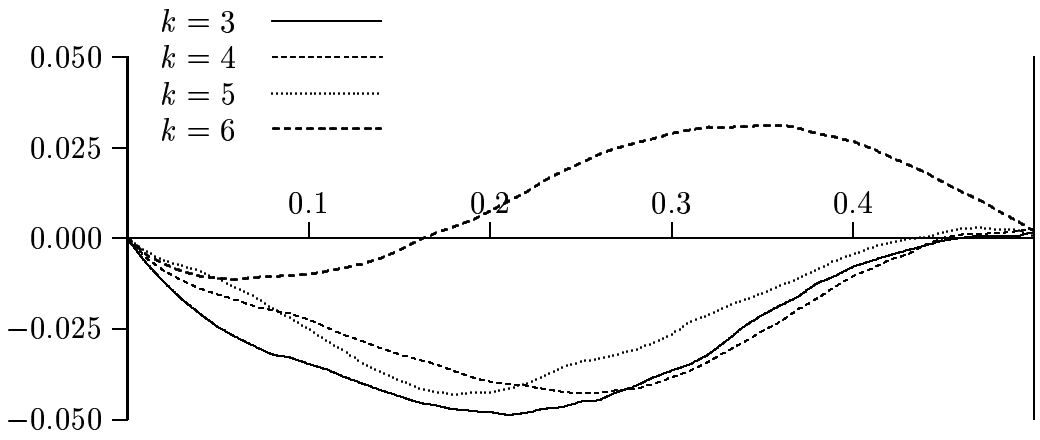

Figure 8a. Base case, test of $\beta_2 = 0$

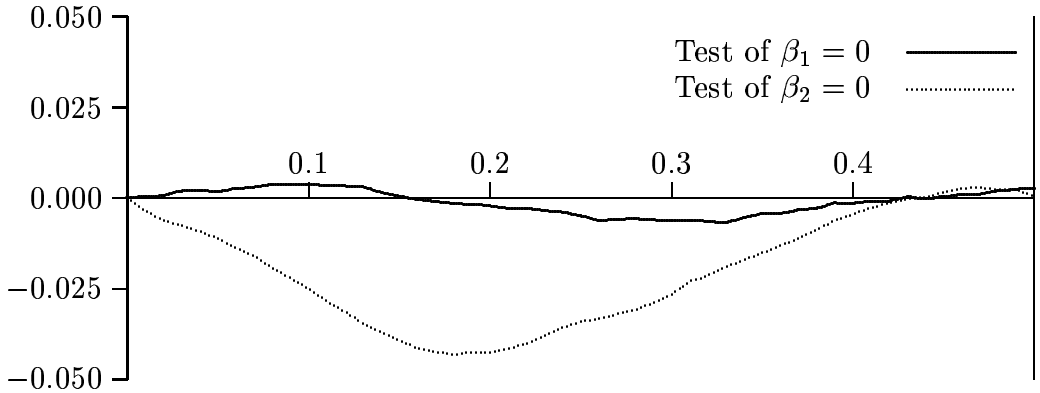

Figure 8b. Tests of $\beta_1 = 0$ and $\beta_2 = 0$ for $k = 5$

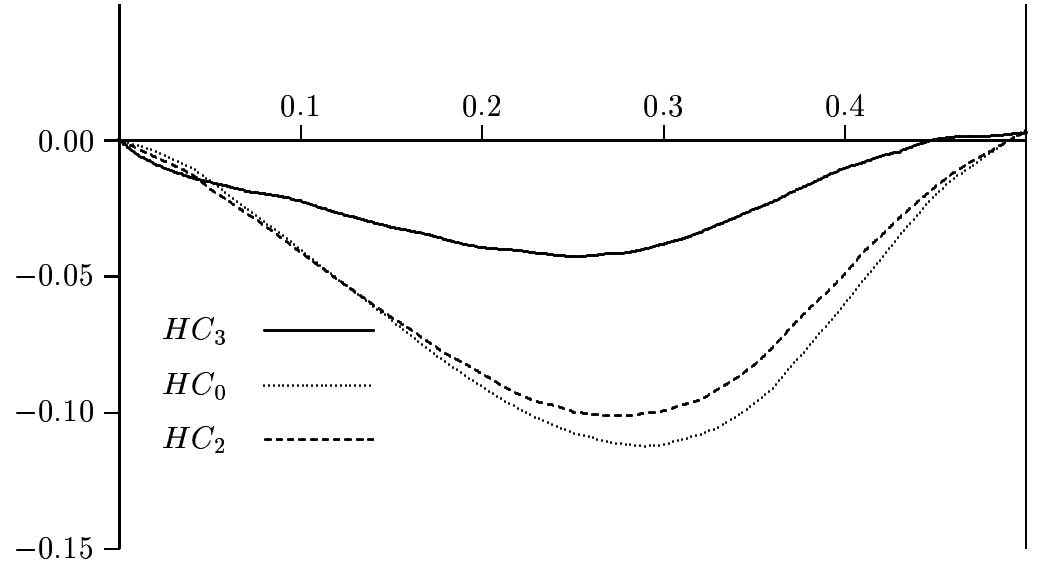

Figure 8c. Tests of $\beta_2 = 0$ for $k = 5$

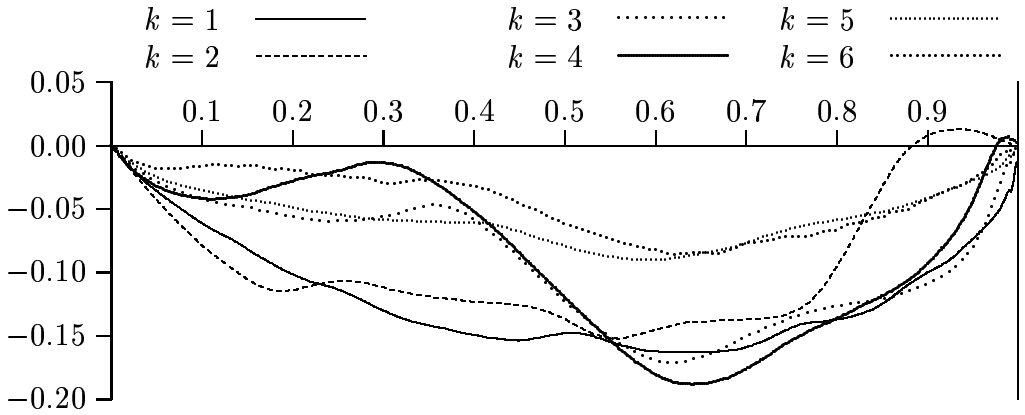

Figure 9a. Skewed error terms

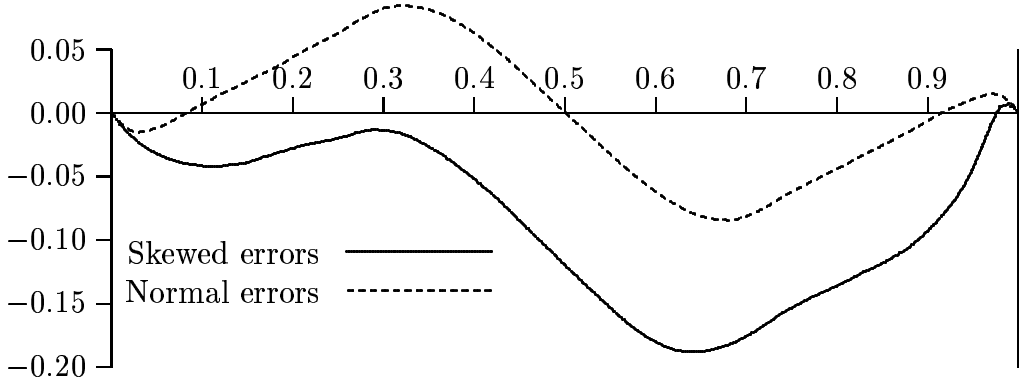

Figure 9b. ERP comparison with normal and skewed errors, $k = 4$

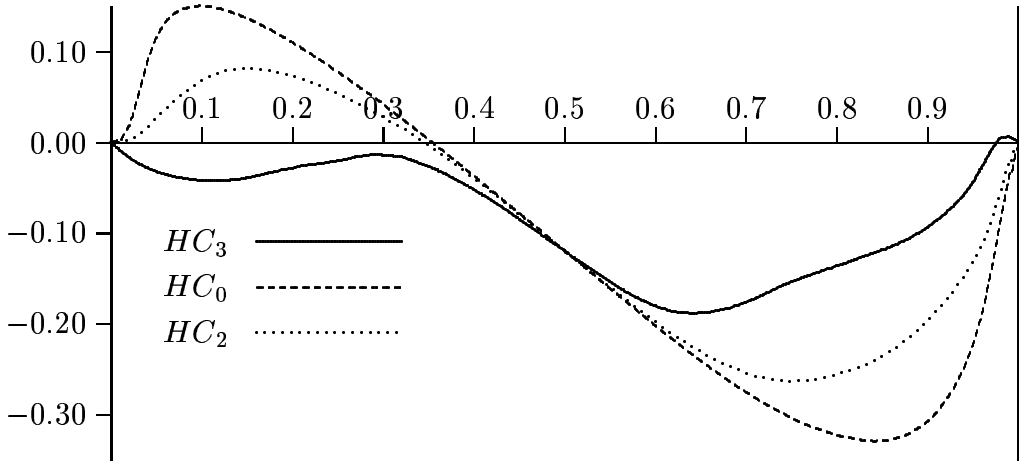

Figure 9c. Skewed errors, HC_0**,** HC_2**, and** HC_3**,** $k = 4$

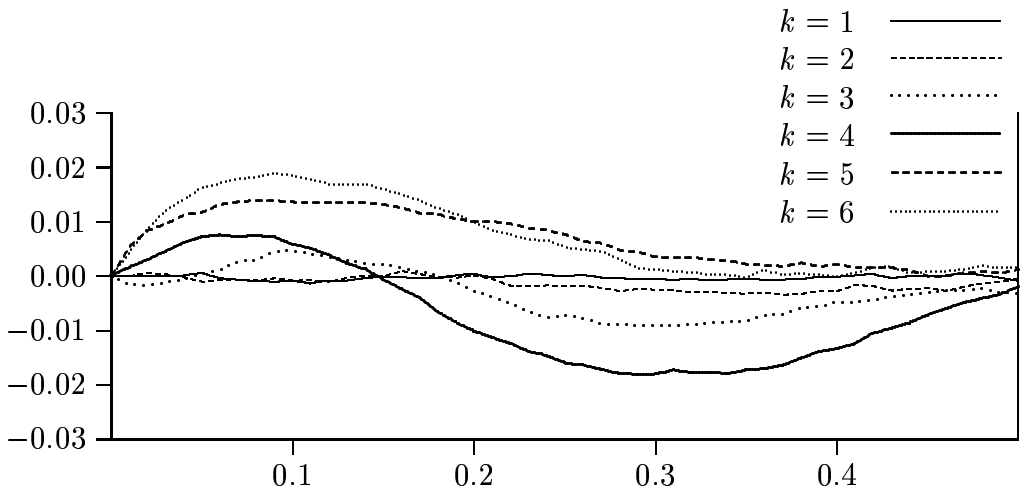

Figure 10a. Homoskedastic normal error terms

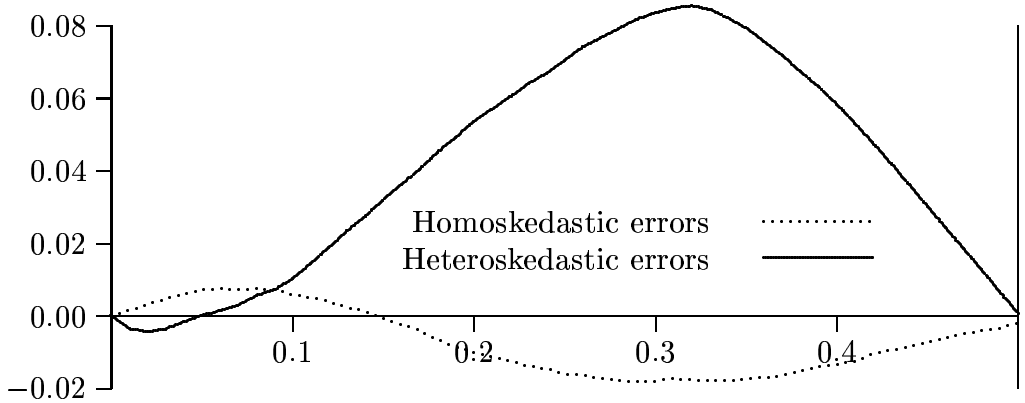

Figure 10b. Comparison homo/heteroskedastic, $k = 4$

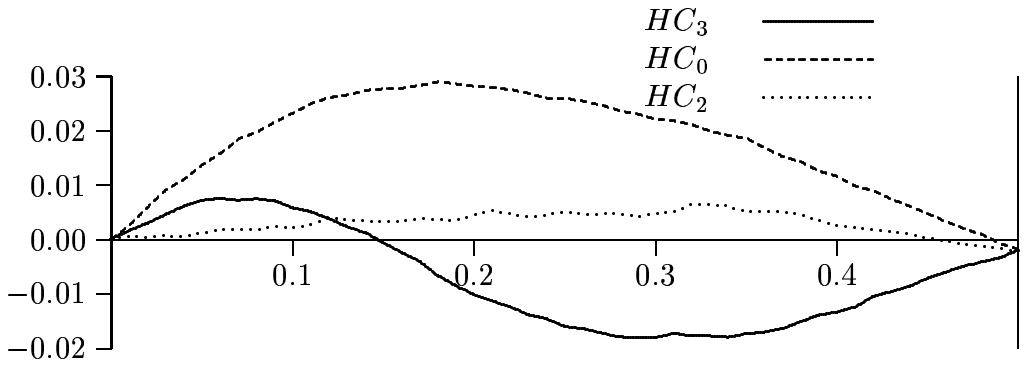

Figure 10c. Homoskedastic errors, HC_0, HC_2, **and** HC_3, $k = 4$

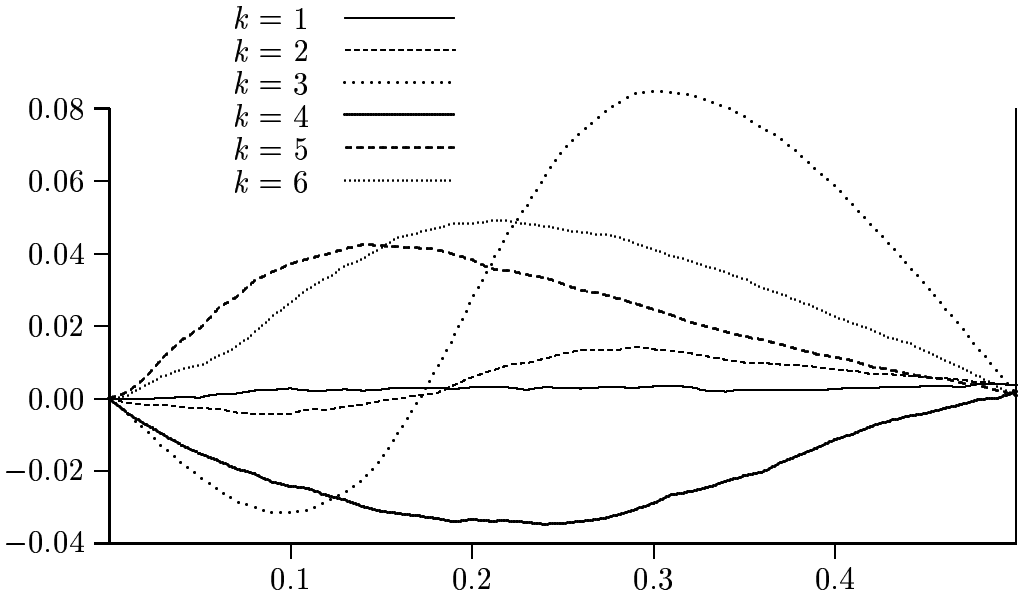

Figure 11a. Balanced design, heteroskedastic errors

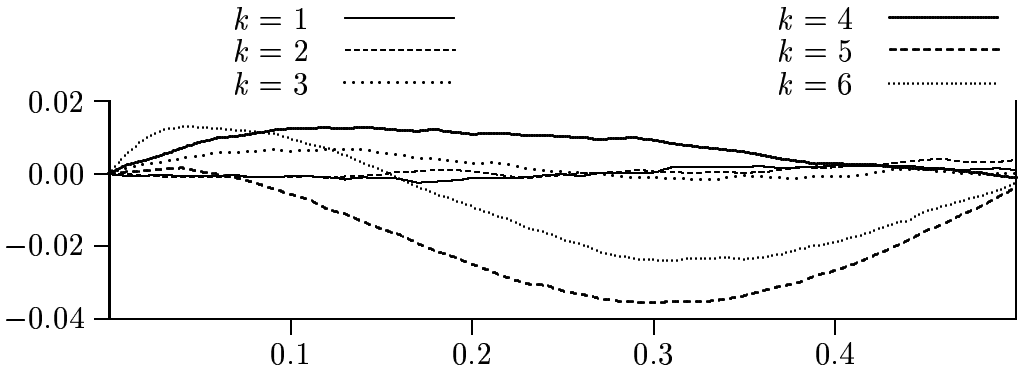

Figure 11b. Balanced design, homoskedastic errors

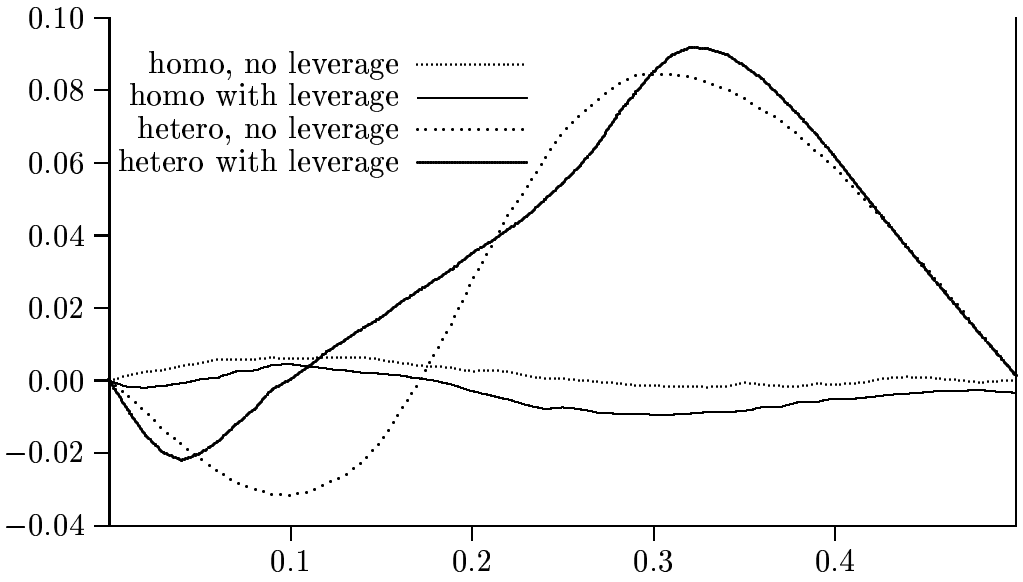

Figure 11c. Leverage/balanced, hetero/homoskedastic, HC_3**,** $k = 3$

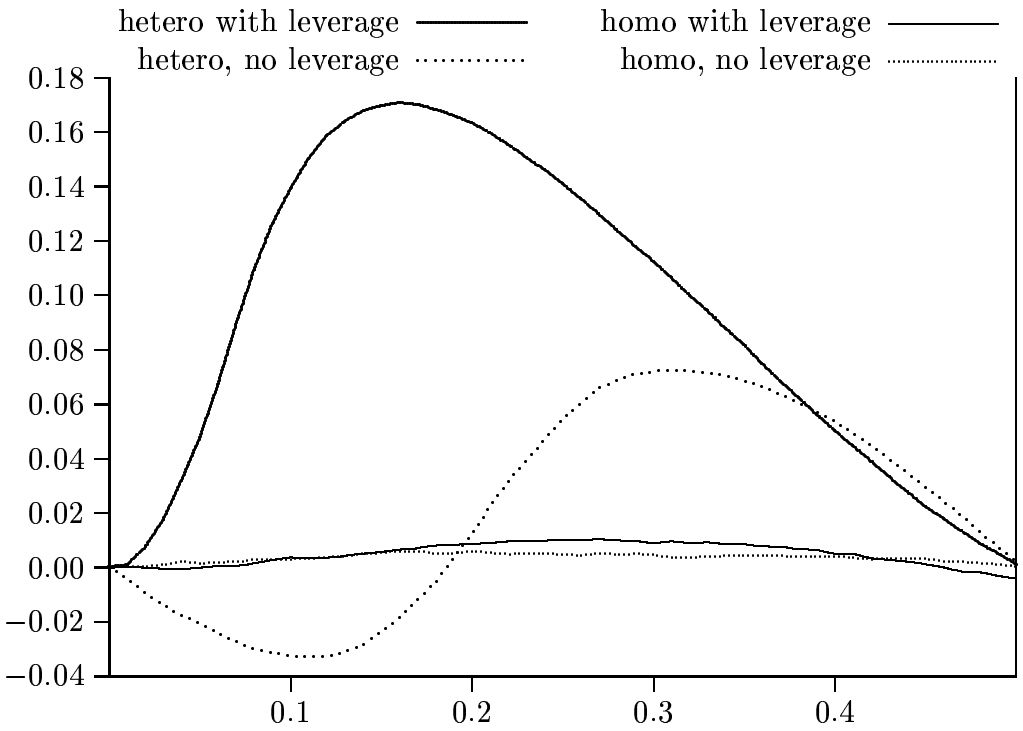

Figure 11d. Leverage/balanced, hetero/homoskedastic, HC_0**,** $k = 3$

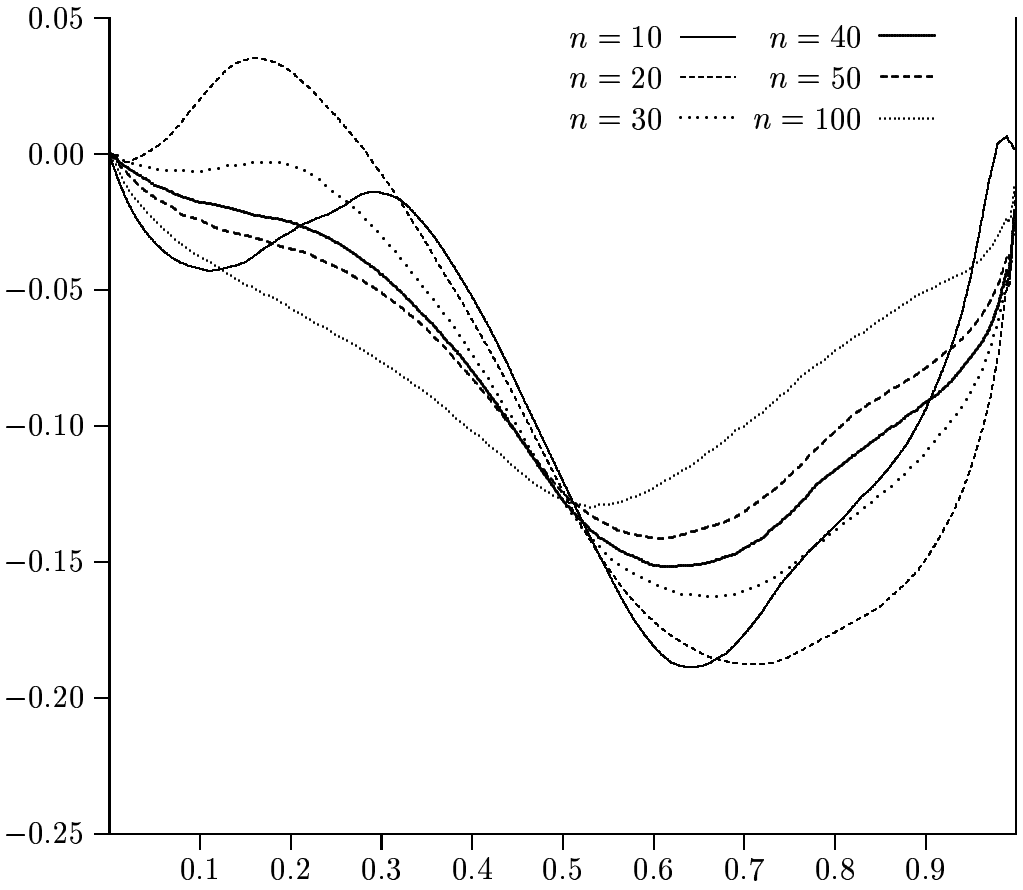

Figure 12a. Skewed errors with different sample sizes

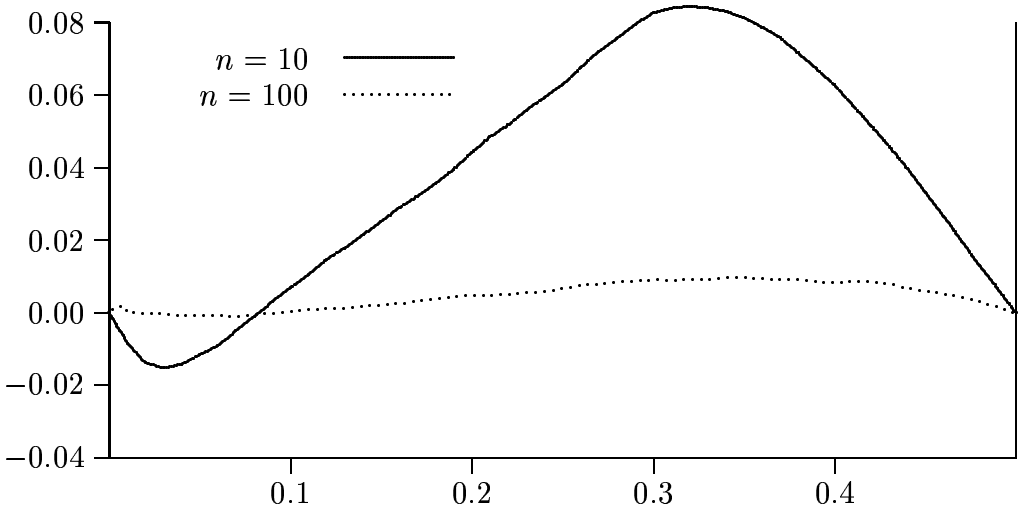

Figure 12b. Normal errors

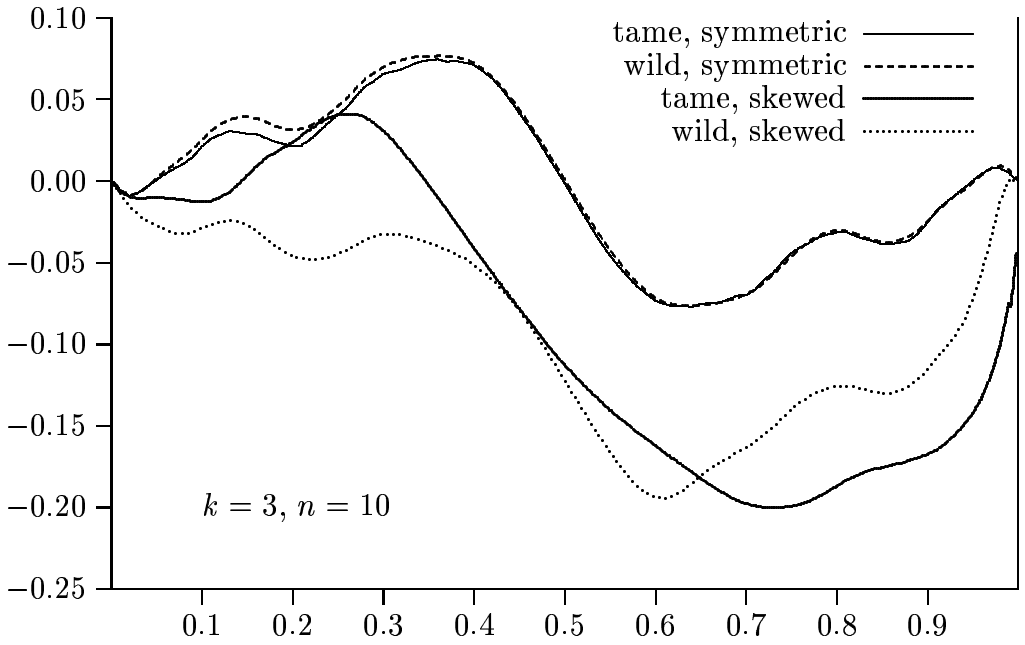

Figure 13a. F_1**-based bootstrap, wild or tamed, symmetric or skewed errors**

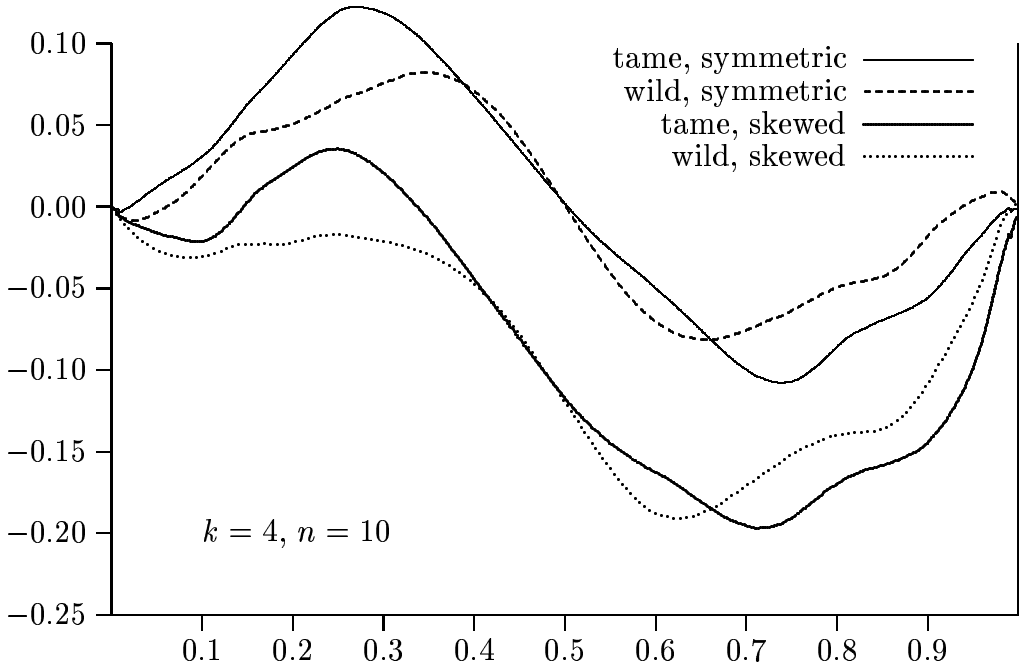

Figure 13b. F_1**-based bootstrap, wild or tamed, symmetric or skewed errors**

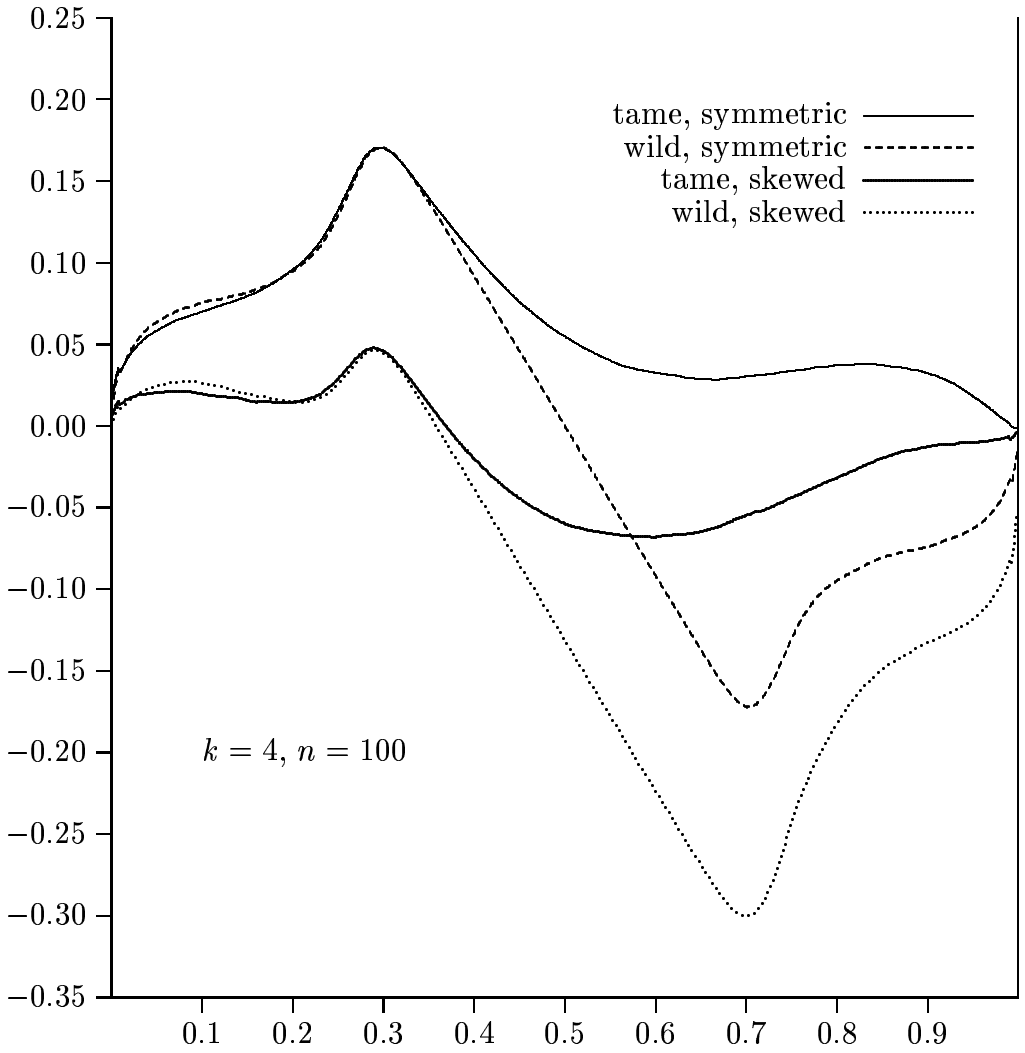

Figure 13c. F_1**-based bootstrap, wild or tamed, symmetric or skewed errors**

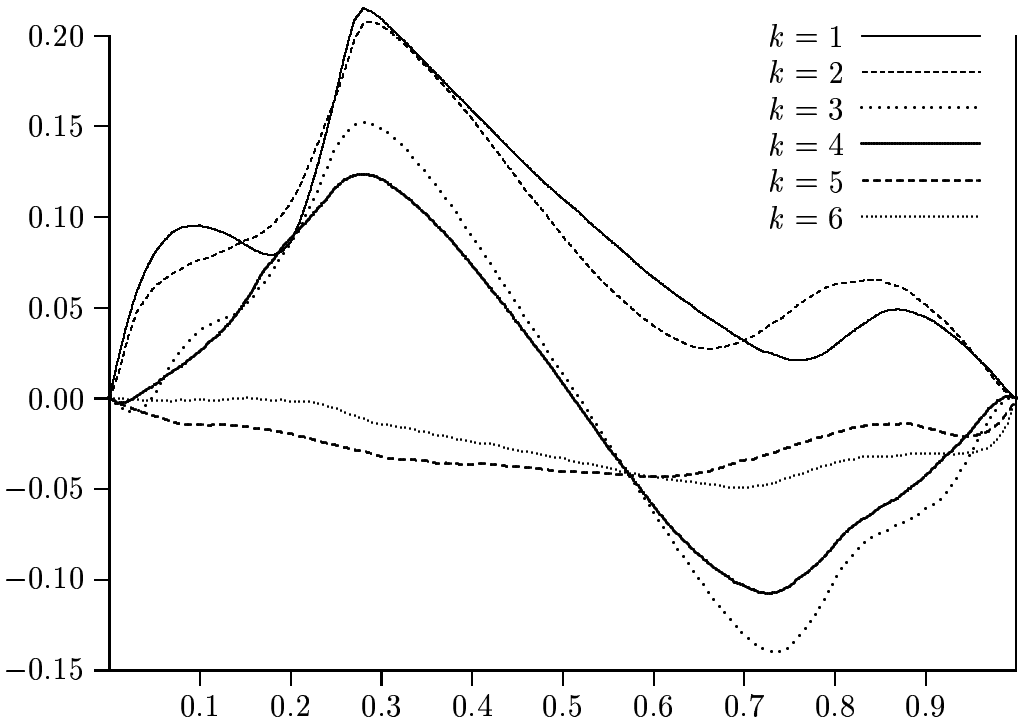

Figure 14a. Base case but with tamed F_1-based bootstrap

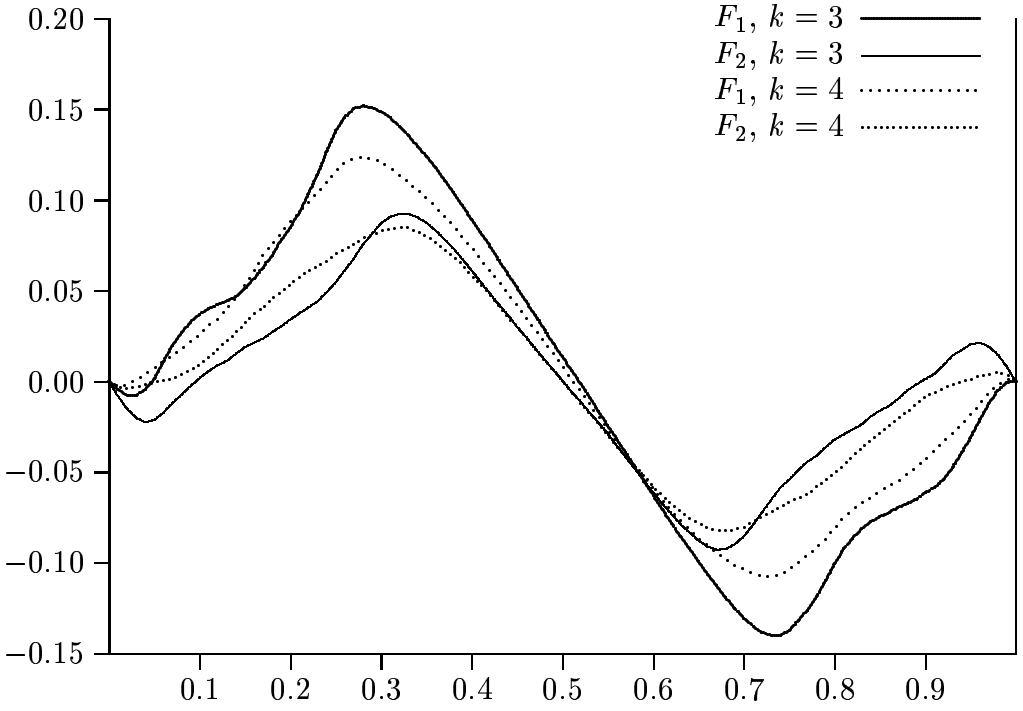

Figure 14b. F_1 and F_2-based bootstraps, normal errors

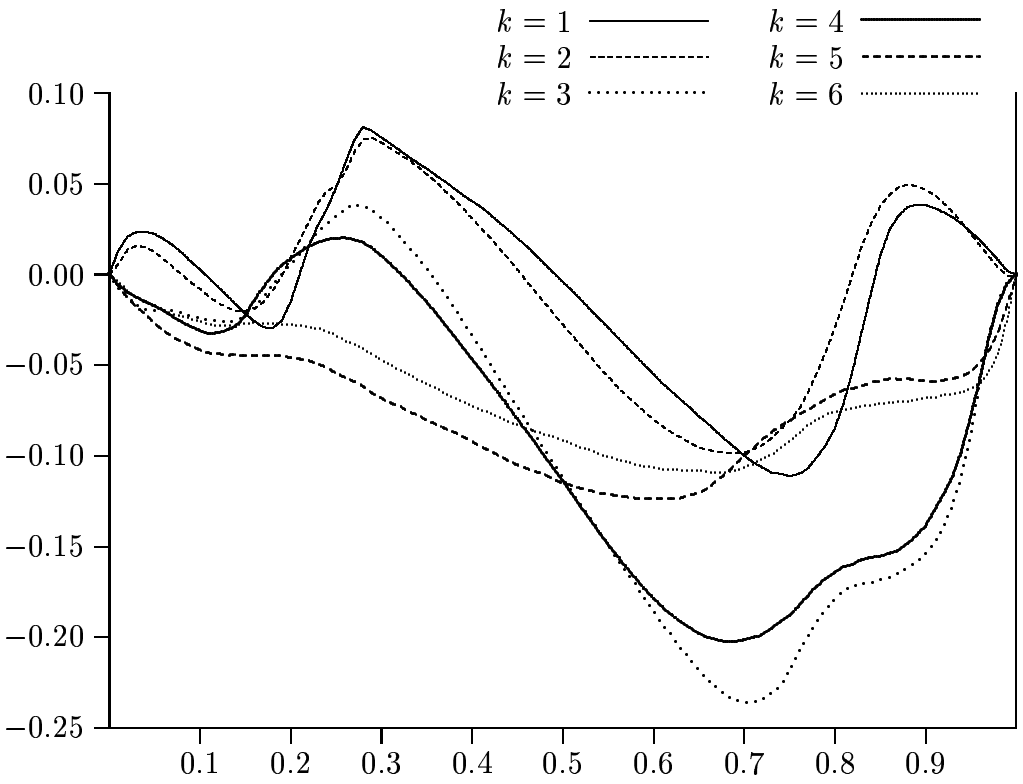

Figure 14c. Skewed errors with tamed F_1-based bootstrap

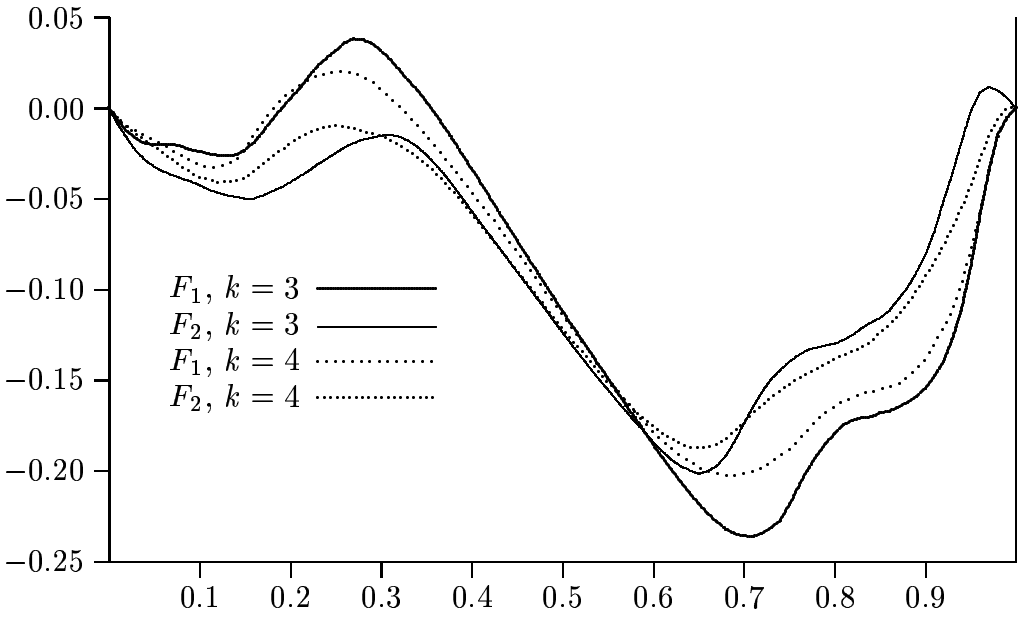

Figure 14d. F_1 and F_2-based bootstraps, skewed errors