Thirty Years of Heteroskedasticity-Robust Inference

James G. MacKinnon

Abstract White (*Econometrica*, 48:817–838, 1980) marked the beginning of a new era for inference in econometrics. It introduced the revolutionary idea of inference that is robust to heteroskedasticity of unknown form, an idea that was very soon extended to other forms of robust inference and also led to many new estimation methods. This paper discusses the development of heteroskedasticity-robust inference since 1980. There have been two principal lines of investigation. One approach has been to modify White's original estimator to improve its finite-sample properties, and the other has been to use bootstrap methods. The relation between these two approaches, and some ways in which they may be combined, are discussed. Finally, a simulation experiment compares various methods and shows how far heteroskedasticity-robust inference has come in just over 30 years.

1 Introduction

White (1980), which appears to be the most cited paper in economics, ushered in a new era for inference in econometrics. The defining feature of this new era is that the distributional assumptions needed for asymptotically valid inference are no longer the same as the ones needed for fully efficient asymptotic inference. The latter still requires quite strong assumptions about disturbances, but the former generally requires much weaker assumptions. In particular, for many econometric models, valid inference is possible in the presence of heteroskedasticity of unknown form, and it

Research for this paper was supported, in part, by a grant from the Social Sciences and Humanities Research Council of Canada. I am grateful to Dimitris Politis, Patrik Guggenberger, and an anonymous referee for comments.

J. G. MacKinnon (✉)
Department of Economics, Queen's University, Kingston, ON K7L 3N6, Canada
e-mail: jgm@econ.queensu.ca

X. Chen and N. R. Swanson (eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, DOI: 10.1007/978-1-4614-1653-1_17,

is often possible as well in the presence of various types of unknown dependence, such as serial correlation and clustered disturbances.

The linear regression model dealt with in White (1980) can be written as

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + u_i, \quad i = 1, \ldots, n, \tag{1}$$

where the $1 \times k$ vectors of regressors $\boldsymbol{X}_i$ may be fixed or random, the disturbances u_i are independent but, in general, not identically distributed, with unknown variances σ_i^2 that may depend on the $\boldsymbol{X}_i$, and certain regularity conditions must be imposed on the pairs $(\boldsymbol{X}_i, u_i)$. The paper proved a number of important asymptotic results, of which the key one is that

$$\hat{\boldsymbol{V}}_n \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{X}_i^\top \boldsymbol{X}_i \xrightarrow{\text{a.s.}} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(u_i^2 \boldsymbol{X}_i^\top \boldsymbol{X}_i), \tag{2}$$

where $\hat{u}_i \equiv y_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ is the ith OLS residual.

In 1980, this was a startling result. The rightmost quantity in (2) is an average of n matrix expectations, and each of those expectations is unknown and impossible to estimate consistently. For many decades, despite a few precursors in the statistics literature such as Eicker (1963, 1967) and Hinkley (1977), econometricians believed that it is necessary to estimate each expectation separately in order to estimate an average of expectations consistently. The key contribution of White (1980) was to show that it is not necessary at all.

The result (2) makes it easy to obtain the asymptotic covariance matrix estimator

$$(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\hat{\boldsymbol{V}}_n(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}, \tag{3}$$

and it is shown in White (1980) that (3) consistently estimates the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$. As the author remarks in a masterpiece of understatement, “This result fills a substantial gap in the econometrics literature, and should be useful in a wide variety of applications.”

The finite-sample covariance matrix estimator that corresponds to (3) is

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\Big(\sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{X}_i^\top \boldsymbol{X}_i\Big)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \tag{4}$$

in which the factors of n have been removed. This estimator, which later came to be known as HC0, was the first *heteroskedasticity-consistent covariance matrix estimator*, or *HCCME*, in econometrics. Estimators that look like (4) are generally referred to as *sandwich covariance matrix estimators*.

Although White (1980) uses the notation $\sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{X}_i^\top \boldsymbol{X}_i$ to denote the filling of the sandwich, most discussions of HCCMEs use the notation $\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{X}$ instead, where $\hat{\boldsymbol{\Omega}}$ is an $n \times n$ diagonal matrix with typical diagonal element $\hat{u}_i^2$. The latter notation is certainly more compact, and I will make use of it in the rest of the paper. However, the more compact notation has two disadvantages. It tends to obscure the fundamental result (2), and it can lead to very inefficient computer programs if they are written in a naive way, because it involves an $n \times n$ matrix.

The key result that averages of expectations can be estimated consistently even when individual ones cannot has had profound implications for econometric theory and practice. It did not take long for econometricians to realize that, if heteroskedasticity-robust inference is possible, then so must be inference that is robust to both heteroskedasticity and autocorrelation of unknown form. Key early papers on what has come to be known as *HAC estimation* include Hansen (1982), White and Domowitz (1984), Newey and West (1987, 1994), Andrews (1991), and Andrews and Monahan (1992). New estimation methods, notably the generalized method of moments (Hansen 1982) and its many variants and offshoots, which would not have been possible without HCCMEs and HAC estimators, were rapidly developed following the publication of White (1980). There were also many important theoretical developments, including White (1982), the key paper on misspecified models in econometrics.

This paper discusses the progress in heteroskedasticity-robust inference since White (1980). Section 2 deals with various methods of heteroskedasticity-consistent covariance matrix estimation. Section 3 deals with bootstrap methods both as an alternative to HCCMEs and as a way of obtaining more reliable inferences based on HCCMEs. Section 4 briefly discusses robust inference for data that are clustered as well as heteroskedastic. Section 5 presents simulation results on the finite-sample properties of some of the methods discussed in Sects. 2 and 3, and the paper concludes in Sect. 6.

2 Better HCCMEs

The HC0 estimator given in expression (4) is not the only finite-sample covariance matrix estimator that corresponds to the asymptotic estimator (3). The matrix (4) depends on squared OLS residuals. Since OLS residuals are on average too small, it seems very likely that (4) will underestimate the true covariance matrix when the sample size is not large. The easiest way to improve (4) is to multiply it by $n/(n-k)$, or, equivalently, to replace the OLS residuals by ones that have been multiplied by $\sqrt{n/(n-k)}$. This is analogous to dividing the sum of squared residuals by $n-k$ instead of by n when we estimate the error variance. This estimator was called HC1 in MacKinnon and White (1985).

MacKinnon and White (1985) also discussed two more interesting procedures. The first of these, which they called HC2 and was inspired by Horn et al. (1975), involves replacing the squared OLS residuals $\hat{u}_i^2$ in (4) by

$$\grave{u}_i^2 \equiv \hat{u}_i^2/(1-h_i),$$

where h_i is the ith diagonal element of the projection matrix $\boldsymbol{P}_{\boldsymbol{X}} \equiv \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, which is sometimes called the *hat matrix*. Because $\mathrm{E}(\hat{u}_i^2) = (1 - h_i)\sigma^2$ when the disturbances are homoskedastic with variance σ^2, it is easy to see that HC2 will be unbiased in that case. In contrast, for HC1 to be unbiased under homoskedasticity, the experimental design must be balanced, which requires that $h_i = k/n$ for all i, a very special case indeed.

The final procedure discussed in MacKinnon and White (1985) was based on the jackknife. In principal, the jackknife involves estimating the model n additional times, each time dropping one observation, and then using the variation among the delete-1 estimates that result to estimate the covariance matrix of the original estimate. For the model (1), this procedure was shown to yield the (finite-sample) estimator

$$\frac{n-1}{n}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\Big(\sum_{i=1}^{n} \acute{u}_i^2 \boldsymbol{X}_i^\top\boldsymbol{X}_i - \frac{1}{n}\boldsymbol{X}^\top\acute{\boldsymbol{u}}\acute{\boldsymbol{u}}^\top\boldsymbol{X}\Big)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \tag{5}$$

where the vector $\acute{\boldsymbol{u}}$ has the typical element

$$\acute{u}_i = \hat{u}_i/(1 - h_i).$$

Notice that, since $\grave{u}_i$ is unbiased when the disturbances are homoskedastic, $\acute{u}_i$ must actually be biased upwards in that case, since $\acute{u}_i = \grave{u}_i/(1 - h_i)^{1/2}$, and the denominator here is always less than one.

MacKinnon and White (1985) called the jackknife estimator (5) HC3, and that is how it is referred to in much of the literature. However, Davidson and MacKinnon (1993) observed that the first term inside the large parentheses in (5) will generally be much larger than the second, because the former is $O_p(n)$ and the latter $O_p(1)$. They therefore (perhaps somewhat cavalierly) redefined HC3 to be the covariance matrix estimator

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\Big(\sum_{i=1}^{n} \acute{u}_i^2 \boldsymbol{X}_i^\top\boldsymbol{X}_i\Big)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \tag{6}$$

which has exactly the same form as HC0, HC1 when the individual OLS residuals are rescaled, and HC2. The modern literature has generally followed this naming convention, and so I will refer to (6) as HC3 and to (5) as HCJ.

Another member of this series of estimators was proposed in Cribari-Neto (2004). The HC4 estimator uses

$$\ddot{u}_i^2 = \hat{u}_i^2/(1 - h_i)^{\delta_i}, \quad \delta_i = \min(4, nh_i/k),$$

instead of the $\hat{u}_i^2$ in (4). The idea is to inflate the ith residual more (less) when h_i is large (small) relative to the average of the h_i, which is k/n. Cribari-Neto and Lima (2009) provide simulation results which suggest that, for the set of models they study, the coverage of confidence intervals based on HCj for $j = 0, \ldots, 4$ always increases monotonically with j. However, HC4 actually overcovers in some cases, so it is not always better than HC3. Poirier (2010) provides an interpretation of HC2 through HC4 in terms of the Bayesian bootstrap. There is also an HC5 estimator, which is quite similar to HC4; see Cribari-Neto et al. (2007).

All of the HCj series of estimators simply modify the (squared) residuals in various ways, but a few papers have taken different approaches. Furno (1996) uses residuals based on robust regression instead of OLS residuals in order to minimize the impact of data points with high leverage. Qian and Wang (2001) and Cribari-Neto and Lima (2010) explicitly correct the biases of various HCCMEs in the HCj series. The formulae that result generally appear to be complicated and perhaps expensive to program when n is large. Both papers present evidence that bias-corrected HCCMEs do indeed reduce bias effectively. However, there appears to be no evidence that they perform particularly well in terms of either coverage for confidence intervals or rejection frequencies for tests. Since those are the things that matter in practice, and bias-corrected HCCMEs are more complicated to program than any of the HCj series, there does not seem to be a strong case for employing the former in applied work.

The relative performance of test statistics and confidence intervals based on different HCCMEs depends principally on the h_i, which determine the leverage of the various observations, and on the pattern of heteroskedasticity. There are valuable analytical results in Chesher and Jewitt (1987), Chesher (1989), and Chesher and Austin (1991). When the sample is balanced, with no points of high leverage, these papers find that HC1, HC2, and HCJ all tend to work quite well. But even a single point of high leverage, especially if the associated disturbance has a large variance, can greatly distort the distributions of test statistics based on some or all of these estimators. Thus, it may be useful to see whether the largest value of h_i is unusually large.

The papers just cited make it clear that HCJ is not always to be preferred to HC2, or even to HC1. In some cases, tests based on HCJ can underreject, and confidence intervals can overcover. The results for HCJ must surely apply to HC3 as well. Similar arguments probably apply with even more force to HC4, which inflates some of the residuals much more than HC3 does; see Sect. 5.

3 Bootstrap Methods

There are two widely used methods for bootstrapping regression models with independent but possibly heteroskedastic disturbances. Both methods can be used to estimate covariance matrices, but they do so in ways that are computationally inefficient and have no theoretical advantages over much simpler methods like HC2 and HC3. In most cases, this is not very useful. What is much more useful is to combine these bootstrap methods with statistics constructed using HCCMEs in order to obtain more reliable inferences than the latter can provide by themselves.

The oldest of the two methods is the *pairs bootstrap* (Freedman 1981), in which the investigator resamples from the entire data matrix. For a linear regression model, or any other model where the data matrix can be expressed as $[\boldsymbol{y}\ \ \boldsymbol{X}]$, each bootstrap sample $[\boldsymbol{y}^* \; \boldsymbol{X}^*]$ simply consists of n randomly chosen rows of the data matrix. We can write a typical bootstrap sample as

$$[\boldsymbol{y}^* \;\; \boldsymbol{X}^*] = \begin{bmatrix} y_{1^*} & \boldsymbol{X}_{1^*} \\ y_{2^*} & \boldsymbol{X}_{2^*} \\ \vdots & \vdots \\ y_{n^*} & \boldsymbol{X}_{n^*} \end{bmatrix},$$

where each of the indices 1^* through n^*, which are different for each bootstrap sample, takes the values 1 through n with probability $1/n$. Thus if, for example, $1^* = 27$ for a particular bootstrap sample, the first row of the data matrix for that sample will consist of the 27th row of the actual data matrix. Technically, the pairs bootstrap data are drawn from the empirical distribution function, or EDF, of the actual data. This is similar to bootstrap resampling for a single variable as originally proposed in Efron (1979, 1982).

Since the regressor matrix will be different for each of the bootstrap samples, the pairs bootstrap does not make sense if the regressors are thought of as fixed in repeated samples. Moreover, to the extent that the finite-sample properties of estimators or test statistics depend on a particular $\boldsymbol{X}$ matrix, the pairs bootstrap may not mimic these properties as well as we would hope because it does not condition on $\boldsymbol{X}$. The pairs bootstrap as just described does not impose any restrictions. However, a modified version for regression models that does allow one to impose restrictions on the bootstrap DGP was proposed in Flachaire (1999).

The original idea of bootstrapping was to estimate standard errors, or more generally the covariance matrices of estimates of parameter vectors, by using the variation among the estimates from the bootstrap samples. If $\hat{\boldsymbol{\beta}}_j^*$ denotes the estimate of $\boldsymbol{\beta}$ from the jth bootstrap sample and $\bar{\boldsymbol{\beta}}^*$ denotes the average of the $\hat{\boldsymbol{\beta}}_j^*$ over B bootstrap samples, the bootstrap estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}$ is simply

$$\widehat{\mathrm{Var}}^*(\hat{\boldsymbol{\beta}}) = \frac{1}{B-1}\sum_{j=1}^{B}(\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*)(\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*)^\top. \tag{7}$$

Although bootstrap covariance matrix estimators like (7) can be useful in some cases (for example, complicated nonlinear models or nonlinear functions of the coefficient estimates in regression models), the matrix (7) is actually just another HCCME, and not one that has any particular merit in finite samples. In fact, Lancaster (2006) shows that the covariance matrix of a delta method approximation to the distribution of the $\hat{\boldsymbol{\beta}}_j^*$ is simply HC0. In practice, when B is large enough, the matrix (7) is probably somewhat better than HC0, but no better than HC2 or HC3.

The main advantage of the pairs bootstrap is that it can be used with a very wide variety of models. For regression models, however, what is generally acknowledged to be a better way to deal with heteroskedasticity is the *wild bootstrap*, which was proposed in Liu (1988) and further developed in Mammen (1993). For the model (1) with no restrictions, the wild bootstrap DGP is

$$y_i^* = \boldsymbol{X}_i \hat{\boldsymbol{\beta}} + f(\hat{u}_i) v_i^*, \tag{8}$$

where $f(\hat{u}_i)$ is a transformation of the ith residual $\hat{u}_i$, and v_i^* is a random variable with mean 0 and variance 1. A natural choice for the transformation $f(\cdot)$ is

$$f(\hat{u}_i) = \frac{\hat{u}_i}{(1-h_i)^{1/2}}. \tag{9}$$

Since this is the same transformation used by HC2, we will refer to it as w2. Using (9) ensures that the $f(\hat{u}_i)$ must have constant variance whenever the disturbances are homoskedastic. Alternatively, one could divide $\hat{u}_i$ by $1-h_i$, which is the transformation that we will refer to as w3 because it is used by HC3. The fact that v_i^* has mean 0 ensures that $f(\hat{u}_i)v_i^*$ also has mean 0, even though $f(\hat{u}_i)$ may not.

Transformations very similar to w2 and w3 can also be useful in the context of bootstrap prediction with homoskedastic errors, where the bootstrap DGP resamples from the rescaled residuals. Stine (1985) suggested using what is essentially w2, and Politis (2010) has recently shown that using predictive (or jackknife) residuals, which effectively use w3, works better.

There are, in principle, many ways to specify the random variable v_i^*. The most popular is the two-point distribution

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

This distribution was suggested in Mammen (1993). Its theoretical advantage is that the skewness of the bootstrap error terms is the same as the skewness of the residuals. A simpler two-point distribution, called the *Rademacher distribution*, is just

$$F_2: \quad v_i^* = \begin{cases} -1 & \text{with probability } \frac{1}{2}, \\ 1 & \text{with probability } \frac{1}{2}. \end{cases}$$

This distribution imposes symmetry on the bootstrap error terms, which it is good to do if they actually are symmetric.

In some respects, the error terms for the wild bootstrap DGP (8) do not resemble those of the model (1) at all. When a two-point distribution like F_1 or F_2 is used, the bootstrap error term can take on only two possible values for each observation. Nevertheless, the wild bootstrap mimics the essential features of the true DGP well enough for it to be useful in many cases.

For any bootstrap method,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^* &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}_j^* - \bar{\boldsymbol{\beta}}^* \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{u}_j^*) - \bar{\boldsymbol{\beta}}^* \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}_j^* + (\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^*), \end{aligned} \tag{10}$$

where $\boldsymbol{y}_j^*$ and $\boldsymbol{u}_j^*$ denote, respectively, the regressand and the vector of error terms for the jth bootstrap sample. If we use the wild bootstrap DGP (8), and the OLS estimator is unbiased, then the expectation of the bootstrap estimates $\hat{\boldsymbol{\beta}}_j^*$ will just be $\hat{\boldsymbol{\beta}}$, and so the last term in the last line of (10) should be zero on average.

The first term in the last line of (10) times itself transposed looks like a sandwich covariance matrix, but with $\boldsymbol{u}_j^*\boldsymbol{u}_j^{*\top}$ instead of a diagonal matrix:

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}_j^*\boldsymbol{u}_j^{*\top}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.$$

It is particularly easy to see what this implies when the bootstrap errors are generated by F_2. In that case, the diagonal elements of $\boldsymbol{u}_j^*\boldsymbol{u}_j^{*\top}$ are simply the squares of the $f(\hat{u}_i)$. The off-diagonal elements must have expectation zero, because, for each bootstrap sample, every off-diagonal element is a product of the same two transformed residuals multiplied either by $+1$ or -1, each with probability one-half. Thus, as B becomes large, we would expect the average of the $\boldsymbol{u}_j^*\boldsymbol{u}_j^{*\top}$ to converge to a diagonal matrix with the squares of the $f(\hat{u}_i)$ on the diagonal. It follows that, if the transformation $f(\cdot)$ is either w2 or w3, the bootstrap covariance matrix estimator (7) must converge to either HC2 or HC3 as $B \to \infty$.

So far, we have seen only that the pairs bootstrap and the wild bootstrap provide computationally expensive ways to approximate various HCCMEs. If that was all these bootstrap methods were good for, there would be no point using them, at least not in the context of making inferences about the coefficients of linear regression models. They might still be useful for calculating covariance matrices for nonlinear functions of those coefficients.

Where these methods, especially the wild bootstrap, come into their own is when they are used together with heteroskedasticity-robust test statistics in order to obtain more accurate P values or confidence intervals. There is a great deal of evidence that the wild bootstrap outperforms the pairs bootstrap in these contexts; see Horowitz (2001), MacKinnon (2002), Flachaire (2005), and Davidson and Flachaire (2008), among others. Therefore, only the wild bootstrap will be discussed.

Consider the heteroskedasticity-robust t statistic

$$\tau(\hat{\beta}_l - \beta_l^0) = \frac{\hat{\beta}_l - \beta_l^0}{\sqrt{\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]_{ll}}}, \tag{11}$$

in which the difference between $\hat{\beta}_l$, the OLS estimate of the lth element of $\boldsymbol{\beta}$ in (1) and its hypothesized value β_l^0 is divided by the square root of the lth diagonal element of any suitable HCCME, such as HC2, HC3, or HC4, depending on precisely how $\hat{\boldsymbol{\Omega}}$ is defined. This test statistic is asymptotically distributed as N(0, 1) under quite weak assumptions. But its finite-sample distribution may or may not be well approximated by the standard normal distribution. Because (11) is asymptotically pivotal, bootstrap methods should provide an asymptotic refinement, that is, more rapid convergence as the sample size increases.

To calculate a wild bootstrap P value for the test statistic (11), we first estimate the model (1) under the null hypothesis to obtain restricted estimates $\tilde{\boldsymbol{\beta}}$ and restricted residuals $\tilde{\boldsymbol{u}}$. We then generate B bootstrap samples, using the DGP

$$y_i^* = \boldsymbol{X}_i \tilde{\boldsymbol{\beta}} + f(\tilde{u}_i) v_i^*. \tag{12}$$

As in (8), there are several choices for the transformation $f(\cdot)$. We have already defined w2 in Eq. (9) and w3 just afterwards. Another possibility, which we will call w1, is just $\sqrt{(n/(n-k+1)}\,\tilde{u}_i$. The random variates v_i^* could be drawn from F_1, F_2, or possibly some other distribution with mean 0 and variance 1.

For each bootstrap sample, indexed as usual by j, we calculate $\tau_j^*(\beta_l)$, the bootstrap analog of the test statistic (11), which is

$$\tau_j^*(\hat{\beta}_{lj}^* - \beta_l^0) = \frac{\hat{\beta}_{lj}^* - \beta_l^0}{\sqrt{\big[(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \hat{\boldsymbol{\Omega}}_j^* \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\big]_{ll}}}. \tag{13}$$

Here, $\hat{\beta}_{lj}^*$ is the OLS estimate for the jth bootstrap sample, and $\boldsymbol{X}^\top \hat{\boldsymbol{\Omega}}_j^* \boldsymbol{X}$ is computed in exactly the same way as $\boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X}$ in (11), except that it uses the residuals from the bootstrap regression.

Davidson and Flachaire (2008) have shown, on the basis of both theoretical analysis and simulation experiments, that wild bootstrap tests based on the Rademacher distribution F_2 can be expected to perform better, in finite samples, than ones based on the Mammen distribution F_1, even when the true disturbances are moderately skewed. Some of the results in Sect. 5 strongly support this conclusion.

Especially when one is calculating bootstrap P values for several tests, it is easier to use unrestricted rather than restricted estimates in the bootstrap DGP, because there is no need to estimate any of the restricted models. The bootstrap data are then generated using (8) instead of (12), and the bootstrap t statistics are calculated as $\tau_j^*(\hat{\beta}_{lj}^* - \hat{\beta}^l)$, which means replacing β_l^0 by $\hat{\beta}_l$ on both sides of Eq. (13). This ensures that the bootstrap test statistics are testing a hypothesis which is true for the bootstrap data.

When using studentized statistics like (11) and other statistics that are asymptotically pivotal, it is almost always better to use restricted estimates in the bootstrap DGP, because the DGP is estimated more efficiently when true restrictions are imposed; see Davidson and MacKinnon (1999). However, this is not true for statistics which are not asymptotically pivotal; see Paparoditis and Politis (2005). The advantage of using restricted estimates can be substantial in some cases, as will be seen in Sect. 5.

Once we have computed $\hat{\tau} = \tau(\hat{\beta}_l - \beta_l^0)$ and B instances of τ_j^*, which be either $\tau_j^*(\hat{\beta}_{lj}^* - \beta_l^0)$ or $\tau_j^*(\hat{\beta}_{lj}^* - \hat{\beta}_l)$, the bootstrap P value is simply

$$\hat{p}^*(\hat{\tau}) = 2 \min\left(\frac{1}{B}\sum_{j=1}^{B} \mathrm{I}(\tau_j^* \le \hat{\tau}), \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}(\tau_j^* > \hat{\tau})\right). \tag{14}$$

This is an *equal-tail bootstrap P value*, so called because, for a test at level α, the rejection region is implicitly any value of $\hat{\tau}$ that is either less than the $\alpha/2$ quantile or greater than the $1-\alpha/2$ quantile of the empirical distribution of the τ_j^*. It is desirable to choose B such that $\alpha(B+1)/2$ is an integer; see Racine and MacKinnon (2007).

For t statistics, it is generally safest to use an equal-tail P value like (14) unless there is good reason to believe that the test statistic is symmetrically distributed around zero. For any test that rejects only when the test statistic is in the upper tail, such as a heteroskedasticity-robust F statistic or the absolute value of a heteroskedasticity-robust t statistic, we would instead compute the bootstrap P value as

$$\hat{p}^*(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}(\tau_j^* > \hat{\tau}). \tag{15}$$

In this case, it is desirable to choose B such that $\alpha(B+1)$ is an integer, which must of course be true whenever $\alpha(B+1)/2$ is an integer.

In many cases, we are interested in confidence intervals rather than tests. The most natural way to obtain a bootstrap confidence interval in this context is to use the *studentized bootstrap*, which is sometimes known as the *percentile-t method*. The bootstrap data are generated using the wild bootstrap DGP (8), which does not impose the null hypothesis. Each bootstrap sample is then used to compute a bootstrap test statistic $\tau_j^*(\hat{\beta}_{lj}^* - \hat{\beta}^l)$. These are sorted, and their $\alpha/2$ and $1-\alpha/2$ quantiles obtained, which is particularly easy to do if $\alpha(B+1)/2$ is an integer. If $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$ denote these empirical quantiles, and $s(\hat{\beta}^l)$ denotes the (presumably heteroskedasticity-robust) standard error of $\hat{\beta}^l$, then the studentized bootstrap interval at level α is

$$\left[\hat{\beta}^l - s(\hat{\beta}^l)q_{1-\alpha/2}^*, \quad \hat{\beta}^l - s(\hat{\beta}^l)q_{\alpha/2}^*\right]. \tag{16}$$

As usual, the lower limit of this interval depends on the upper tail quantile of the bootstrap test statistics, and the upper limit depends on the lower tail quantile. Even if the true distribution of the τ_j^* happens to be symmetric around the origin, it is highly unlikely that the empirical distribution will be. Therefore, the interval (16) will almost never be symmetric.

Another way to find confidence intervals is explicitly to invert bootstrap P values. The confidence interval then consists of all points for which the bootstrap P value (14) is greater than α. Solving for such an interval can be a bit complicated, since the null hypotheses that correspond to each end of the interval must be imposed on the bootstrap DGP. However, this technique can be more reliable than the studentized bootstrap method; see Davidson and MacKinnon (2010, 2011).

The discussion so far may have incorrectly given the impression that the only reason to use the wild bootstrap is to reduce the size distortion of tests, or the coverage errors of confidence intervals, that are associated with HCCMEs which are not entirely reliable in finite samples. In cross-section regressions with samples of several hundred observations or more, those errors are often quite modest. But there may well be other sources of much larger size distortions or coverage errors that can also be reduced by using bootstrap methods. Although the primary reason for bootstrapping may not be heteroskedasticity of unknown form, it is often wise to use a technique like the wild bootstrap together with heteroskedasticity-robust covariance matrices.

An important example is two-stage least squares (or generalized IV) estimation with possibly heteroskedastic disturbances when the instruments are not strong. Davidson and MacKinnon (2010) proposed a wild bootstrap procedure for this case. When there are just two endogenous variables, the model is

$$\boldsymbol{y}_1 = \beta \boldsymbol{y}_2 + \boldsymbol{Z}\gamma + \boldsymbol{u}_1 \tag{17}$$

$$\boldsymbol{y}_2 = \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{u}_2. \tag{18}$$

Equation (17) is a structural equation, and Eq. (18) is a reduced-form equation. The n-vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are vectors of observations on endogenous variables, $\boldsymbol{Z}$ is an $n \times k$ matrix of observations on exogenous variables, and $\boldsymbol{W}$ is an $n \times l$ matrix of exogenous instruments with the property that $\mathbb{S}(\boldsymbol{Z})$, the subspace spanned by the columns of $\boldsymbol{Z}$, lies in $\mathbb{S}(\boldsymbol{W})$, the subspace spanned by the columns of $\boldsymbol{W}$. Typical elements of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are denoted by y_{1i} and y_{2i} respectively, and typical rows of $\boldsymbol{Z}$ and $\boldsymbol{W}$ are denoted by $\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$.

Davidson and MacKinnon (2010) discusses several wild bootstrap procedures for testing the hypothesis that $\beta = \beta_0$. The one that works best, which they call the *wild restricted efficient* (or *WRE*) bootstrap, uses the bootstrap DGP

$$y^*_{1i} = \beta_0 y^*_{2i} + \boldsymbol{Z}_i \tilde{\gamma} + f_1(\tilde{u}_{1i}) v^*_i \tag{19}$$

$$y^*_{2i} = \boldsymbol{W}_i \tilde{\boldsymbol{\pi}} + f_2(\tilde{u}_{2i}) v^*_i, \tag{20}$$

where $\tilde{\gamma}$ and the residuals $\tilde{u}_{1i}$ come from an OLS regression of $\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2$ on $\boldsymbol{Z}$, $\tilde{\boldsymbol{\pi}}$ comes from an OLS regression of $\boldsymbol{y}_2$ on $\boldsymbol{W}$ and $\tilde{\boldsymbol{u}}_1$, and $\tilde{\boldsymbol{u}}_2 \equiv \boldsymbol{y}_2 - \boldsymbol{W}\tilde{\boldsymbol{\pi}}$. The transformations $f_1(\cdot)$ and $f_2(\cdot)$ could be any of w1, w2, or w3.

This bootstrap DGP has three important features. First, the structural Eq. (19) uses restricted (OLS) estimates instead of unrestricted (2SLS) ones. This is very important for the finite-sample properties of the bootstrap tests. Note that, if 2SLS estimates were used, it would no longer make sense to transform the $\hat{u}_{1i}$, because 2SLS residuals are not necessarily too small. Second, the parameters of the reduced-form Eq. (20) are estimated efficiently, because the structural residuals are included as an additional regressor. This is also very important for finite-sample properties. Third, the same random variable v_i^* multiplies the transformed residuals for both equations. This ensures that the correlation between the structural and reduced-form residuals is retained by the structural and reduced-form bootstrap error terms.

Davidson and MacKinnon (2010) provides evidence that bootstrap tests of hypotheses about β based on the WRE bootstrap perform remarkably well whenever the sample size is not too small (400 seems to be sufficient) and the instruments are not very weak. What mostly causes asymptotic tests to perform poorly is simultaneity combined with weak instruments, and not heteroskedasticity. The main reason to use the WRE bootstrap is to compensate for the weak instruments.

Ideally, one should always use a heteroskedasticity-robust test statistic together with the wild bootstrap, or perhaps some other bootstrap method that is valid in the presence of heteroskedasticity. However, it is also asymptotically valid to use a nonrobust test statistic together with the wild bootstrap, or a robust test statistic together with a bootstrap method that does not take account of heteroskedasticity. The simulation evidence in Davidson and MacKinnon (2010) suggests that both of these approaches, while inferior to the ideal one, can work reasonably well.

4 Cluster-Robust Covariance Matrices

An important extension of heteroskedasticity-robust inference is cluster-robust inference. Consider the linear regression model

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_m \end{bmatrix} \equiv \boldsymbol{X\beta} + \boldsymbol{u}.$$

Here, there are m clusters, indexed by j, the observations for which are stacked into the vector $\boldsymbol{y}$ and the matrix $\boldsymbol{X}$. Clusters might correspond to cities, counties, states, or countries in a cross-section of households or firms, or they might correspond to cross-sectional units in a panel dataset. The important thing is that there may be correlation among the disturbances within each cluster, but not across clusters.

If we know nothing about the pattern of variance and covariances within each cluster, then it makes sense to use a cluster-robust covariance matrix estimator. The simplest such estimator is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \biggl(\sum_{j=1}^{m} \boldsymbol{X}_j^\top \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^\top \boldsymbol{X}_j \biggr) (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{21}$$

where $\hat{\boldsymbol{u}}_j$ is the vector of OLS residuals for the jth cluster. This has the familiar sandwich form of an HCCME, except that the filling in the sandwich is more complicated. It is robust to heteroskedasticity of unknown form as well as to within-cluster correlation. The estimator (21) was first proposed by Froot (1989), introduced into Stata by Rogers (1993), and extended to allow for serial correlation of unknown form, as in HAC estimation, by Driscoll and Kraay (1998). It is widely used in applied work.

Cameron et al. (2008) recently proposed a wild bootstrap method for clustered data. As in the usual wild bootstrap case, where the bootstrap disturbance for observation i depends on the residual $\hat{u}_i$, all the bootstrap disturbances for each cluster depend on the residuals for that cluster. The wild bootstrap DGP is

$$y^*_{ji} = \boldsymbol{X}_{ji}\hat{\boldsymbol{\beta}} + f(\hat{u}_{ji})v^*_j, \tag{22}$$

where j indexes clusters, i indexes observations within each cluster, and the v^*_j follow the Rademacher (F_2) distribution. The key feature of (22) is that there are only as many v^*_j as there are clusters. Thus, the bootstrap DGP preserves the variances and covariances of the residuals within each cluster. This method apparently works surprisingly well even when the number of clusters is quite small.

5 Simulation Evidence

Simulation experiments can be used to shed light on the finite-sample performance of various HCCMEs, either used directly for asymptotic tests or combined with various forms of the wild bootstrap. This section reports results from a number of experiments that collectively deal with a very large number of methods. Most of the experiments were deliberately designed to make these methods perform poorly.

Many papers that use simulation to study the properties of HCCMEs, beginning with MacKinnon and White (1985) and extending at least to Cribari-Neto and Lima (2010), have simply chosen a fixed or random $\boldsymbol{X}$ matrix for a small sample size—just 10 in the case of Davidson and Flachaire (2008)—and formed larger samples by repeating it as many times as necessary. When $\boldsymbol{X}$ matrices are generated in this way, there will only be as many distinct values of h_i as the number of observations in the original sample. Moreover, all of those values, and in particular the largest one, must be exactly proportional to $1/n$; see Chesher (1989). This ensures that inference based on heteroskedasticity-robust methods improves rapidly as n increases. Since very few real datasets involve $\boldsymbol{X}$ matrices for which all of the h_i are proportional to $1/n$, this sort of experiment almost certainly paints an excessively optimistic picture. Some evidence on this point is provided below.

In contrast, the model employed here, which is similar to one used for a much more limited set of experiments in MacKinnon (2002), is

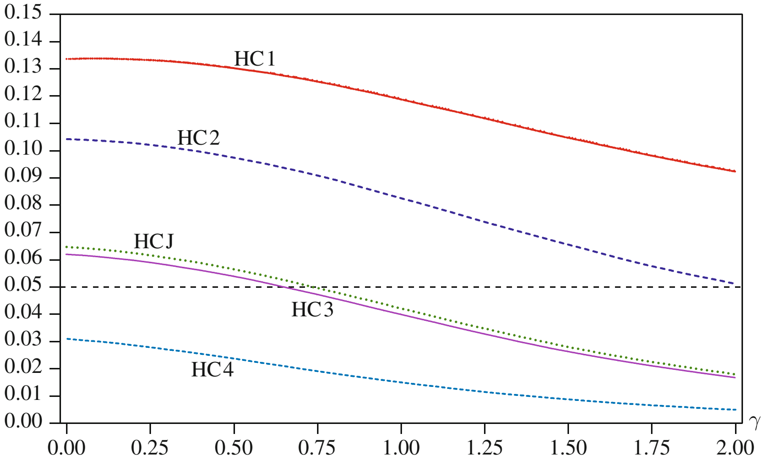

Fig. 1 Rejection frequencies for heteroskedasticity-robust t tests, $n = 40$

$$y_i = \beta_1 + \sum_{k=2}^{5} \beta_k X_{ik} + u_i, \quad u_i = \sigma_i \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, 1), \tag{23}$$

where all regressors are drawn randomly from the standard lognormal distribution, $\beta_k = 1$ for $k \leq 4$, $\beta_5 = 0$, and

$$\sigma_i = z(\gamma)\big(\beta_1 + \sum_{k=2}^{5} \beta_k X_{ik}\big)^{\gamma}. \tag{24}$$

Here, $z(\gamma)$ is a scaling factor chosen to ensure that the average variance of u_i is equal to 1. Thus, changing the parameter γ changes how much heteroskedasticity there is but does not, on average, change the variance of the disturbances. In the experiments, $0 \leq \gamma \leq 2$. Note that $\gamma = 0$ implies homoskedasticity, and $\gamma >> 1$ implies rather extreme heteroskedasticity.

The DGP consisting of Eqs. (23) and (24) was deliberately chosen so as to make heteroskedasticity-robust inference difficult. Because the regressors are lognormal, many samples will contain a few observations on the X_{ik} that are quite extreme, and the most extreme observation in each sample will tend to become more so as the sample size increases. Therefore, the largest value of h_i will tend to be large and to decline very slowly as $n \to \infty$. In fact, the average value of $h_i^{\max}$ is nearly 0.80 when $n = 20$ and declines by a factor of only about 3.5 as the sample size increases to 1,280, with the rate of decline increasing somewhat as n becomes larger. It is likely that few real datasets have h_i which are as badly behaved as the ones in these experiments, so their results almost certainly paint an excessively pessimistic picture.

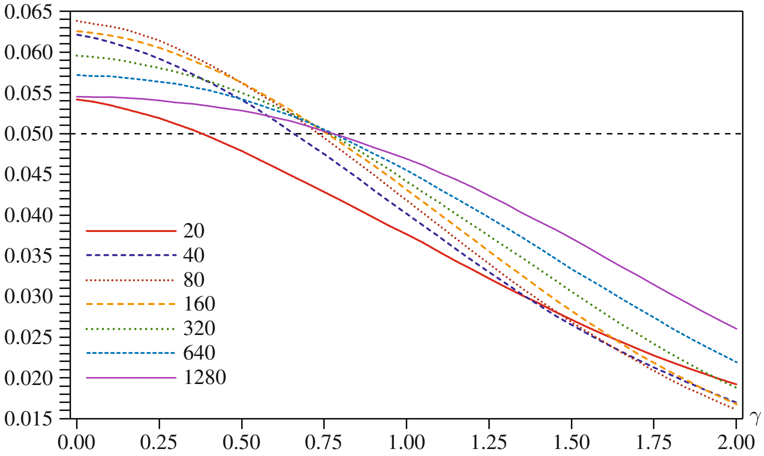

Fig. 2 Rejection frequencies for asymptotic HC3 t tests, various sample sizes

Figures 1 and 2 show the results of several sets of experiments for asymptotic tests of the hypothesis that $\beta_5 = 0$ based on test statistics like (11) and the standard normal distribution. The figures show rejection frequencies as functions of the parameter γ. They are based on 1,000,000 replications for each of 41 values of γ between 0.00 and 2.00 at intervals of 0.05.

Rejection frequencies for five different HCCMEs when $n = 40$ are shown in Fig. 1. As expected, tests based on HC1 always overreject quite severely. Perhaps somewhat unexpectedly, tests based on HC4 always underreject severely. This is presumably a consequence of the very large values of $h_i^{\max}$ in these experiments. Tests based on the other estimators sometimes overreject and sometimes underreject. In every case, rejection frequencies decline monotonically as γ increases. For any given value of γ, they also decline as j increases from 1 to 4 in HCj. It is reassuring to see that the results for HC3 and HCJ are extremely similar, as predicted by Davidson and MacKinnon (1993) when they introduced the former as an approximation to the latter and appropriated its original name.

Note that, as Davidson and Flachaire (2008) emphasized, restricting attention to tests at the 0.05 level is not inconsequential. All the tests are more prone to overreject at the 0.01 level and less prone to overreject at the .10 level than they are at the 0.05 level. In other words, the distributions of the test statistics have much thicker tails than does the standard normal distribution. Even HC4, which underrejects at the 0.05 level for every value of n and γ, always overrejects at the 0.01 level for some small values of γ.

Figure 2 focuses on HC3, which seems to perform best among the HCj estimators for $n = 40$. It shows results for seven values of n from 20 to 1,280. The surprising thing about this figure is how slowly the rejection frequency curves become flatter as the sample size increases. The curves actually become steeper as n increases from 20 to 40 and then to 80. The worst overrejection for $\gamma = 0$ and the worst underrejection

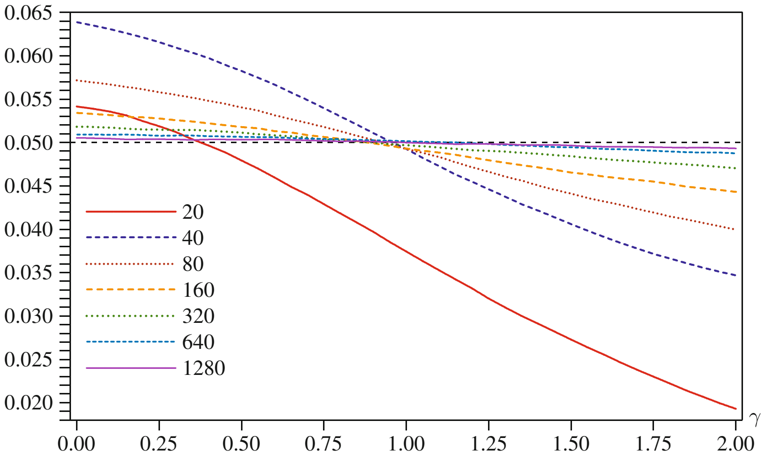

Fig. 3 Rejection frequencies for asymptotic HC3 t tests, 20 rows of $\boldsymbol{X}$ repeated

for $\gamma = 2$ both occur when $n = 80$. As n increases from 80 to 160, 320, 640, and finally 1280, the curves gradually become flatter, but they do so quite slowly. It seems likely that we would need extremely large samples for rejection frequencies to be very close to the nominal level of 0.05 for all values of γ. This is a consequence of the experimental design, which ensures that $h_i^{\max}$ decreases very slowly as n increases.

An alternative to generating the entire regressor matrix for each sample size is simply to generate the first 20 rows and then repeat them as many times as necessary to form larger samples with integer multiples of 20 observations. As noted earlier, $h_i^{\max}$ would then be proportional to $1/n$. Figure 3 contains the same results as Fig. 2, except that the matrix $\boldsymbol{X}$ is generated in this way. The performance of asymptotic tests based on HC3 now improves much faster as n increases. In particular, the rejection frequency curve changes dramatically between $n = 20$ and $n = 40$. It is evident that the way in which $\boldsymbol{X}$ is generated matters enormously.

The remaining figures deal with wild bootstrap tests. Experiments were performed for 12 variants of the wild bootstrap. There are three transformations of the residuals (denoted by w1, w2, or w3, because they are equivalent to HC1, HC2, or HC3), two types of residuals (restricted and unrestricted, denoted by r or u), and two ways of generating the v_j^* (F_1 or F_2, denoted by 1 or 2). The 12 variants are

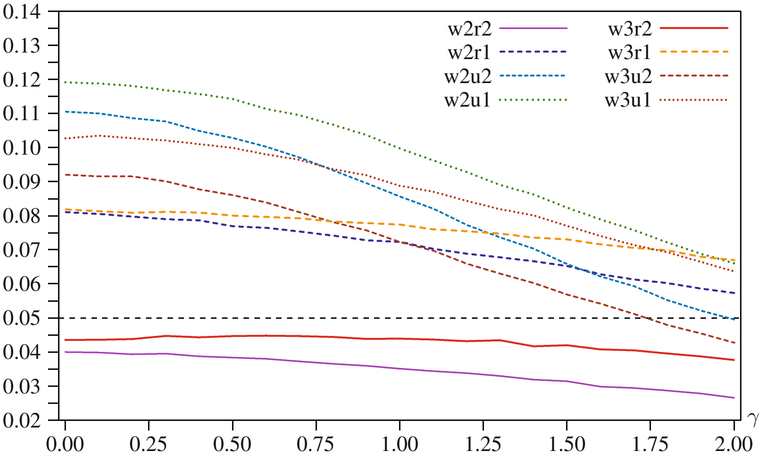

Fig. 4 Rejection frequencies for bootstrap HC3 t tests, $n = 40$

$$\text{w1r1 and w1r2: } u_i^* = \sqrt{n/(n-k+1)}\,\tilde{u}_i$$

$$\text{w1u1 and w1u2: } u_i^* = \sqrt{n/(n-k)}\,\hat{u}_i$$

$$\text{w2r1 and w2r2: } u_i^* = \frac{\tilde{u}_i}{(1-\tilde{h}_i)^{1/2}}$$

$$\text{w2u1 and w2u2: } u_i^* = \frac{\hat{u}_i}{(1-h_i)^{1/2}}$$

$$\text{w3r1 and w3r2: } u_i^* = \frac{\tilde{u}_i}{(1-\tilde{h}_i)}$$

$$\text{w3u1 and w3u2: } u_i^* = \frac{\hat{u}_i}{(1-h_i)}$$

In the expressions for w2r1, w2r2, w3r1, and w3r2, $\tilde{h}_i$ denotes the ith diagonal of the hat matrix for the restricted model.

The experimental results are based on 100,000 replications for each of 21 values of γ between 0.0 and 2.0 at intervals of 0.1, with $B = 399$. In practice, it would be better to use a larger number for B in order to obtain better power, but 399 is adequate in the context of a simulation experiment; see Davidson and MacKinnon (2000). There are five different HCCMEs and 12 different bootstrap DGPs. Thus, each experiment produces 60 sets of rejection frequencies. It would be impossible to present all of these graphically without using an excessively large number of figures.

Figures 4 and 5 present results for HC3 and HC1 respectively, combined with eight different bootstrap DGPs for $n = 40$. Results are shown only for w2 and w3, because the diagram would have been too cluttered if w1 had been included, and methods based on w1 usually performed less well than ones based on w2. HC3 was chosen because asymptotic tests based on it performed best, and HC1 was chosen

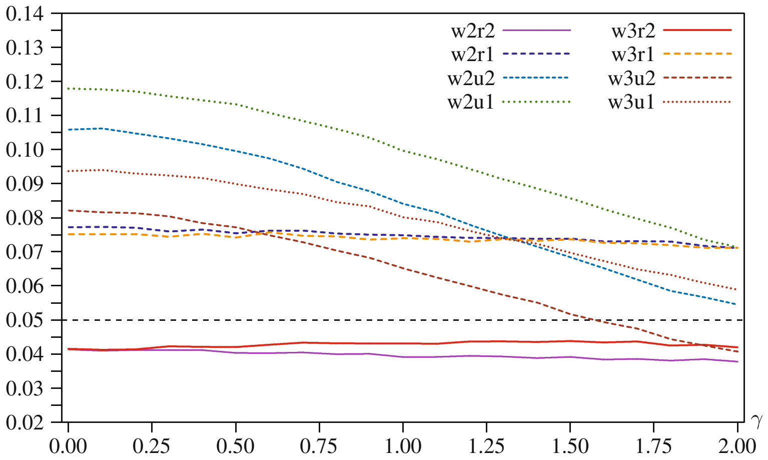

Fig. 5 Rejection frequencies for bootstrap HC1 t tests, $n = 40$

because asymptotic tests based on it performed worst. Note that the results for HC0 would have been identical to the ones for HC1, because the former is just a multiple of the latter. This implies that the position of $\hat{\tau}$ in the sorted list of $\hat{\tau}$ and the τ_j^* must be the same for HC0 and HC1, and hence the P value must be the same.

In Fig. 4, we see that only two of the wild bootstrap methods (w3r2 and w2r2) yield tests that never overreject. The size distortion for w3r2 is always less than for w2r2. The curve for w1r2, not shown, lies everywhere below the curve for w2r2. The other six wild bootstrap methods do not perform particularly well. They all overreject for all or most values of γ. For small values of γ, the four worst methods are the ones that use unrestricted residuals. But w2r1 and w3r1 also work surprisingly poorly.

Figure 5 shows results for the same eight wild bootstrap methods as Fig. 4, but this time the test statistic is based on HC1. The results are similar to those in Fig. 4, but they are noticeably better in several respects. Most importantly, w3r2 and, especially, w2r2 underreject less severely, and all of the tests that use unrestricted residuals overreject somewhat less severely.

The remaining figures focus on the effects of sample size. Figure 6 shows rejection frequencies for tests based on HC1 for six sample sizes, all using the w3r2 wild bootstrap. In striking contrast to the asymptotic results in Fig. 2, the improvement as n increases is quite rapid. Except for the largest values of γ, the rejection frequencies are very close to 0.05 for $n = 640$.

Figure 7 shows that using unrestricted residuals harms performance greatly for all sample sizes. Although there is much faster improvement with n than for the asymptotic tests in Fig. 2, overrejection for small values of γ is actually more severe for the smaller sample sizes. Both overrejection for small values of γ and underrejection for large ones remain quite noticeable even when $n = 640$.

Figure 8 is similar to Fig. 6, except that the matrix $\boldsymbol{X}$ consists of the first 20 rows repeated as many times as necessary. Results are presented only for $n = 40$, 60, 80,

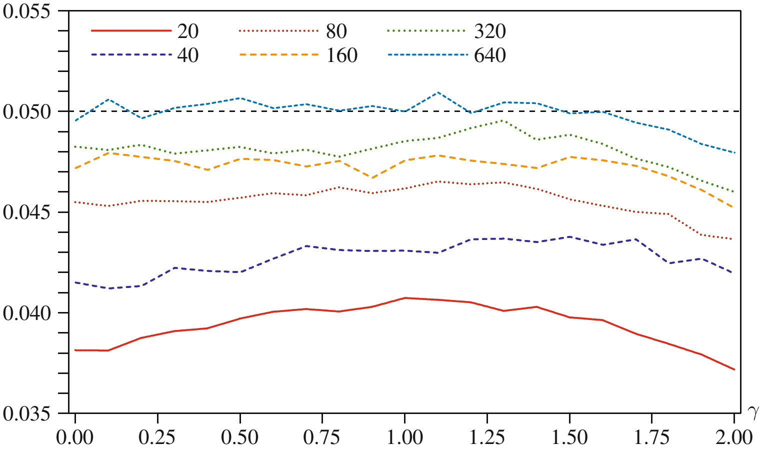

Fig. 6 Rejection frequencies for w3r2 bootstrap HC1 t tests

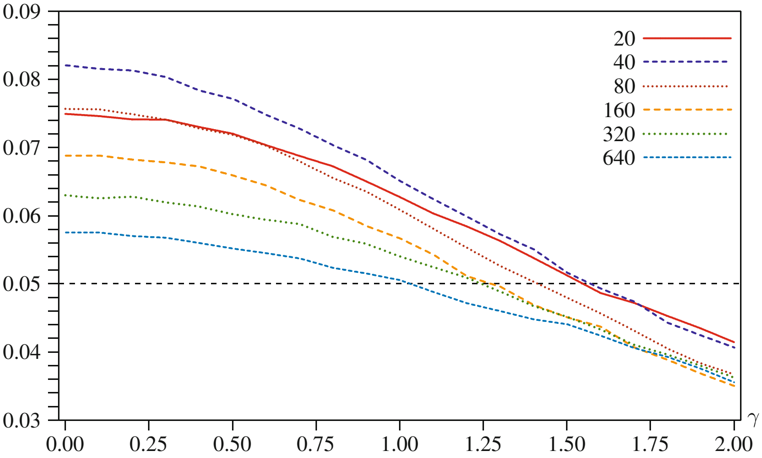

Fig. 7 Rejection frequencies for w3u2 bootstrap HC1 t tests

120, and 160. Results for $n = 20$ are omitted, because they may be found in Fig. 6, and including them would have required a greatly extended vertical axis. Results for sample sizes larger than 160 are omitted for obvious reasons. To reduce experimental error, these results are all based on 400,000 replications.

The performance of all the bootstrap tests in Fig. 8 is extremely good. Simply making the bottom half of the $\boldsymbol{X}$ matrix repeat the top half, as happens when $n = 40$, dramatically improves the rejection frequencies. Results would have been similar for tests based on HC2, HC3, HCJ, or HC4. It is now very difficult to choose among bootstrap tests that use different HCCMEs, as they all work extremely well.

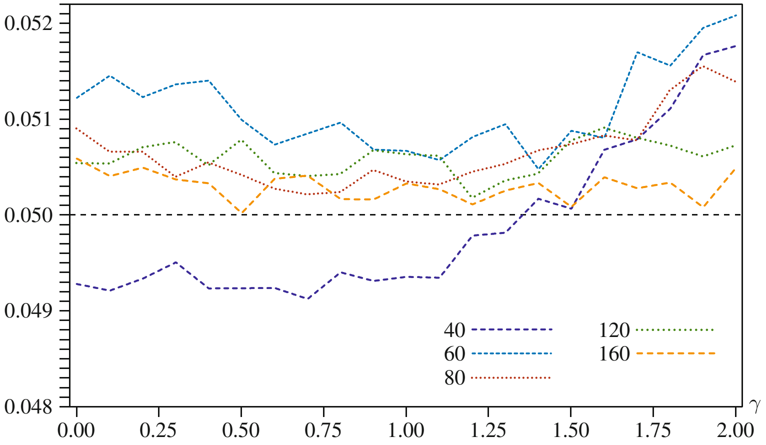

Fig. 8 Rejection frequencies for w3r2 bootstrap HC1 t tests, 20 rows of $\boldsymbol{X}$ repeated

Although Fig. 8 only shows results for the w3r2 variant of the wild bootstrap, other bootstrap methods also perform much better when the regressor matrix consists of the first 20 rows repeated than when the entire matrix is generated randomly. But there still seem to be clear benefits from using restricted residuals and the F_2 distribution, at least for smaller values of n.

Like most of the work in this area, the experiments described so far focus exclusively on the performance of tests under the null. However, test power can be just as important as test size. The remaining experiments therefore address two important questions about power. The first is whether the choice of HCCME matters, and the second is whether there is any advantage to using unrestricted rather than restricted residuals in the bootstrap DGP. These experiments use the w3 bootstrap and the F_2 distribution. The sample size is 40, there are 100,000 replications, and $B = 999$. The number of bootstrap samples is somewhat larger than in the previous experiments, because power loss is proportional to $1/B$; see Jöckel (1986).

Figure 9 shows power functions for wild bootstrap (w3r2) tests of the hypothesis that $\beta_5 = 0$ in Eq. 11 as a function of the actual value of β_5 when $\gamma = 1$. Experiments were performed for 71 values of β_5: -0.70, -0.68, …, 0.68, 0.70. This figure has two striking features. First, the power functions are not symmetric. There is evidently greater power against negative values of β_5 than against positive ones. This occurs because of the pattern of heteroskedasticity in Eq. (24). For $\gamma > 0$, there is more heteroskedasticity when $\beta_5 > 0$ than when $\beta_5 < 0$. This causes the estimate of β_5 to be more variable as the true value of β_5 increases. When $\gamma = 0$, the power functions are symmetric.

The second striking feature of Fig. 9 is that power decreases monotonically from HC1 to HC2, HC3, and finally HC4. Thanks to the bootstrap, all the tests have essentially the same performance under the null. Thus, what we see in the figure is a real, and quite substantial, reduction in power as we move from HC1, which pays

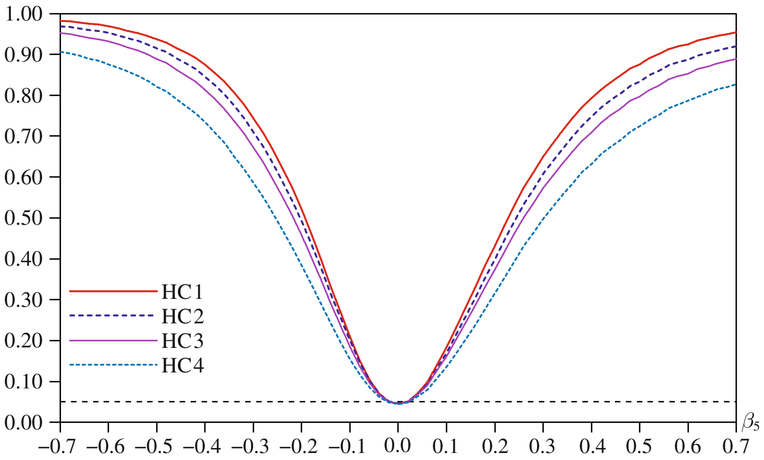

Fig. 9 Power of wild bootstrap (w3r2) heteroskedasticity-robust t tests, $\gamma = 1$, $n = 40$

no attention to the leverage of each observation, to robust covariance matrices that take greater and greater account thereof. At least in this case, there appears to be a real cost to using HCCMEs that compensate for leverage (they overcompensate, in the case of HC3 and HC4). It seems to be much better to correct for the deficiencies of HC1 by bootstrapping rather than by using a different HCCME.

Although there is still the same pattern for similar experiments in which the $\boldsymbol{X}$ matrix is generated by repeating the first 20 observations (results not shown), the loss of power is much less severe. Thus, it may well be that the power loss in Fig. 9 is just about as severe as one is likely to encounter.

It is widely believed that using unrestricted residuals in a bootstrap DGP yields greater power than using restricted residuals. The argument is that, when the null is false, restricted residuals will be larger than unrestricted ones, and so the bootstrap error terms will be too big if restricted residuals are used. Paparoditis and Politis (2005) show that there is indeed a loss of power from using restricted residuals whenever a test statistic is asymptotically nonpivotal. However, their theoretical analysis yields no such result for asymptotically pivotal statistics like the robust t statistic (11) studied here. Using restricted residuals does indeed cause the bootstrap estimates to be more variable, but it also causes the standard errors of those estimates to be larger. Thus, there is no presumption that bootstrap critical values based on the distribution of the bootstrap t statistics will be larger if one uses restricted residuals.

Figure 10 shows two power functions, one for the w3r2 bootstrap, which is identical to the corresponding one in Fig. 9, and one for the w3u2 bootstrap. Using unrestricted residuals causes the test to reject more frequently for most, but not all, values of β_5, including $\beta_5 = 0$. Ideally, one would like to adjust both tests to have precisely the correct size, but this is very difficult to do in a way that is unambiguously correct; see Davidson and MacKinnon (2006). If one could do so, it is not clear that the w3u2 bootstrap would ever have greater power than the w3r2 bootstrap, and it is clear that it would actually have less power for many positive values of β_5.

It can be dangerous to draw conclusions from simulation experiments, especially in a case like this where the details of the experimental design are evidently very important. Nevertheless, it seems to be possible to draw several qualified conclusions from these experiments. Many of these echo previous theoretical and simulation results that may be found in Chesher and Austin (1991), Davidson and Flachaire (2008), and other papers, but others appear to be new.

- The best HCCME for asymptotic inference may not be the best one for bootstrap inference.
- When regressor matrices of various sizes are created by repeating a small number of observations as many times as necessary, both asymptotic and bootstrap tests perform better than they do when there is no repetition and $h_i^{\max}$ decreases slowly as n increases.
- Rejection frequencies for bootstrap tests can improve much more rapidly as n increases than ones for asymptotic tests, even when $h_i^{\max}$ decreases very slowly as n increases.
- Although well-chosen bootstrap methods can work much better than purely asymptotic ones, not all bootstrap methods work particularly well when $h_i^{\max}$ decreases slowly as n increases.
- There can be a substantial gain from using restricted residuals in the wild bootstrap DGP, especially when $h_i^{\max}$ decreases slowly as n increases.
- There can be a substantial gain from using F_2 rather than F_1 to generate the bootstrap error terms, especially when $h_i^{\max}$ decreases slowly as n increases.
- The power of bootstrap tests based on different HCCMEs can differ substantially. The limited evidence presented here suggests that HC1 may yield the greatest power and HC4 the least.
- There is no theoretical basis for, and no evidence to support, the idea that using unrestricted residuals in the bootstrap DGP will yield a more powerful test than using restricted residuals when the test statistic is asymptotically pivotal.

All the experiments focused on testing rather than confidence intervals. However, studentized bootstrap confidence intervals like (16) simply involve inverting bootstrap t tests based on unrestricted residuals. Thus, the poor performance of tests that use unrestricted residuals in the bootstrap DGP when $h_i^{\max}$ decreases slowly suggests that studentized bootstrap confidence intervals may not be particularly reliable when the data have that feature. In such cases, it is likely that one can obtain bootstrap confidence intervals with much better coverage by inverting bootstrap P values based on restricted residuals; see Davidson and MacKinnon (2011).

The base case for these experiments, in which the regressors are randomly generated from the log-normal distribution, is probably unrealistic. In practice, the performance of heteroskedasticity-robust tests may rarely be as bad for moderate and large sample sizes as it is in these experiments. But the other case, in which the rows of the regressor matrix repeat themselves every 20 observations, is even more

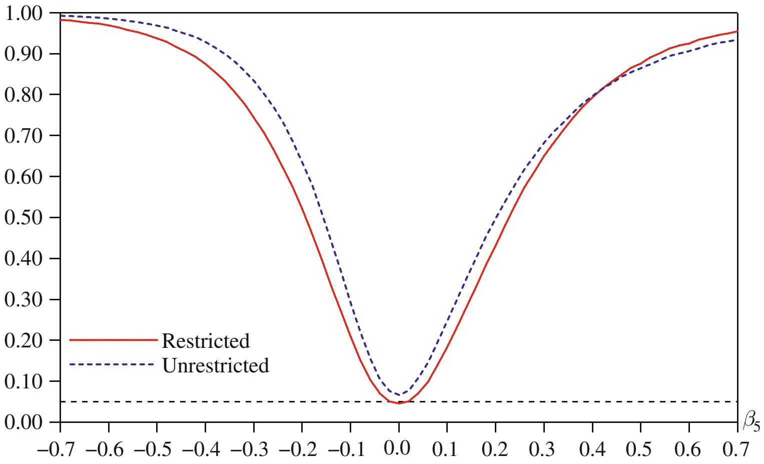

Fig. 10 Power of wild bootstrap HC1 t tests, $\gamma = 1$, $n = 40$

unrealistic. The many published simulation results that rely on this type of experimental design are almost certainly much too optimistic in their assessments of how well heteroskedasticity-robust tests and confidence intervals perform.

6 Conclusion

White (1980) showed econometricians how to make asymptotically valid inferences in the presence of heteroskedasticity of unknown form, and the impact of that paper on both econometric theory and empirical work has been enormous. Two strands of later research have investigated ways to make more accurate inferences in samples of moderate size. One strand has concentrated on finding improved covariance matrix estimators, and the other has focused on bootstrap methods. The wild bootstrap is currently the technique of choice. It has several variants, some of which are closely related to various HCCMEs. The wild bootstrap is not actually a substitute for a good covariance matrix estimator. Instead, it should be used in conjunction with one to provide more accurate tests and confidence intervals. This paper has discussed both strands of research and presented simulation results on the finite-sample performance of asymptotic and bootstrap tests.

Acknowledgments Research for this paper was supported, in part, by a grant from the Social Sciences and Humanities Research Council of Canada. I am grateful to Dimitris Politis, Patrik Guggenberger, and an anonymous referee for comments.

References

Andrews, D.W.K. (1991). “Heteroskedasticity and autocorrelation consistent covariance matrix estimation”, Econometrica, 59:817–858.

Andrews, D.W.K., and J. C. Monahan (1992). “An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator”, Econometrica, 60, 953–966.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). “Bootstrap-based improvements for inference with clustered errors”, Review of Economics and Statistics, 90, 414–427.

Chesher, A. (1989). “Hájek inequalities, measures of leverage and the size of heteroskedasticity robust tests”, Journal of Econometrics, 57, 971–977.

Chesher, A., and G. Austin (1991). “The finite-sample distributions of heteroskedasticity robust Wald statistics”, Journal of Econometrics, 47, 153–173.

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

Cribari-Neto, F. (2004). “Asymptotic inference under heteroskedasticity of unknown form”, Computational Statistics and Data Analysis, 45, 215–233.

Cribari-Neto, F., and M. G. A. Lima (2009). “Heteroskedasticity-consistent interval estimators”, Journal of Statistical Computation and Simulation, 79, 787–803.

Cribari-Neto, F., and M. G. A. Lima (2010). “Sequences of bias-adjusted covariance matrix estimators under heteroskedasticity of unknown form”, Annals of the Institute of Mathematical Statistics, 62, 1053–1082.

Cribari-Neto, F., T. C. Souza, and K. L. P. Vasconcellos (2007). “Inference under heteroskedasticity and leveraged data”, Communications in Statistics: Theory and Methods, 36, 1977–1988 [see also Erratum (2008), 37, 3329–3330.].

Davidson, R., and E. Flachaire (2008). “The wild bootstrap, tamed at last”, Journal of Econometrics, 146, 162–169.

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

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

Davidson, R., and J. G. MacKinnon (2000). “Bootstrap tests: How many bootstraps?” Econometric Reviews, 19, 55–68.

Davidson, R., and J. G. MacKinnon (2006). “The power of bootstrap and asymptotic tests”, Journal of Econometrics, 133, 421–441.

Davidson, R., and J. G. MacKinnon (2010). “Wild bootstrap tests for IV regression”, Journal of Business and Economic Statistics, 28, 128–144.

Davidson, R., and J. G. MacKinnon (2011). “Confidence sets based on inverting Anderson-Rubin tests”, Queen’s University, QED Working Paper No. 1257.

Driscoll, J. C., and A. C. Kraay (1998). “Consistent covariance matrix estimation with spatially dependent panel data”, Review of Economics and Statistics, 80, 549–560.

Efron, B. (1979). “Bootstrap methods: Another look at the jackknife”, Annals of Statistics, 7, 1–26.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia, Society for Industrial and Applied Mathematics.

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

Eicker, F. (1967). “Limit theorems for regressions with unequal and dependent errors”, in Fifth Berkeley Symposium on Mathematical Statistics and Probability, ed. L. M. Le Cam and J. Neyman, Berkeley, University of California Press, 1, 59–82.

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

Flachaire, E. (2005). “Bootstrapping heteroskedastic regression models: Wild bootstrap vs pairs bootstrap”, Computational Statistics and Data Analysis, 49, 361–376.

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

Froot, K. A. (1989). "Consistent covariance matrix estimation with cross-sectional dependence and heteroskedasticity in financial data", Journal of Financial and Quantitative Analysis, 24, 333–355.

Furno, M. (1996). "Small sample behavior of a robust heteroskedasticity consistent covariance matrix estimator", Journal of Statistical Computation and Simulation, 54, 115–128.

Hansen, L. P. (1982). "Large sample properties of generalized method of moments estimators", Econometrica, 50, 1029–1054.

Hinkley, D.V. (1977). "Jackknifing in unbalanced situations", Technometrics, 19, 285–292.

Horn, S. D., R. A. Horn, and D. B. Duncan (1975). "Estimating heteroskedastic variances in linear models", Journal of the American Statistical Association, 70, 380–385.

Horowitz, J. L. (2001). "The bootstrap", in Handbook of Econometrics, Vol. 5, ed. J. J. Heckman and E. E. Leamer. Amsterdam, North-Holland, 3159–3228.

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

Lancaster, T. (2006). "A note on bootstraps and robustness", Brown University Working Paper 2006–6.

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

MacKinnon, J. G. (2002). "Bootstrap inference in econometrics", Canadian Journal of Economics, 35, 615–645.

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.

Newey, W.K., and K. D. West (1987). "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix", Econometrica, 55, 703–708.

Newey, W. K., and K. D. West (1994). "Automatic lag selection in covariance matrix estimation", Review of Economic Studies, 61,631–653.

Paparoditis, E., and D. N. Politis (2005). "Bootstrap hypothesis testing in regression models", Statistics and Probability Letters, 74, 356–365.

Poirier, D. J. (2010). "Bayesian interpretations of heteroskedastic consistent covariance estimators using the informed Bayesian bootstrap", Econometric Reviews, 30, 457–468.

Politis, D.N. (2010). "Model-free model-fitting and predictive distributions", Department of Economics, Univ. of California-San Diego.

Qian, L., and S. Wang (2001). "Bias-corrected heteroscedasticity robust covariance matrix (sandwich) estimators", Journal of Statistical Computation and Simulation, 70, 161–174.

Racine, J. S., and J. G. MacKinnon (2007). "Simulation-based tests that can use any number of simulations", Communications in Statistics: Simulation and Computation, 36, 357–365.

Rogers, W. H. (1993). "Regression standard errors in clustered samples", STATA Technical Bulletin, 13, 19–23.

Stine, R. A. (1985). "Bootstrap prediction intervals for regression", Journal of the American Statistical Association, 80, 1026–1031.

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

White, H. (1982). "Maximum likelihood estimation of misspecified models", Econometrica, 50, 1–25.

White, H., and I. Domowitz (1984). "Nonlinear regression with dependent observations", Econometrica, 52, 143–161.