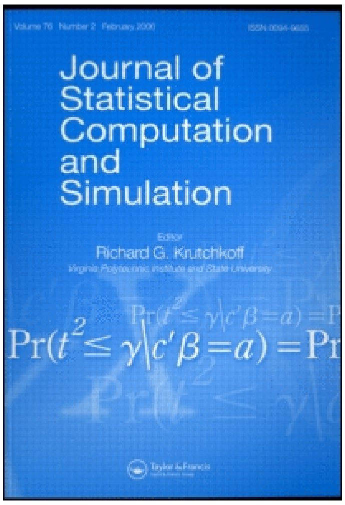

Journal of Statistical Computation and Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713650378

Accurately sized test statistics with misspecified conditional homoskedasticity

Jack Erb[a]; Douglas G. Steigerwald[a]
[a] Department of Economics, University of California, Santa Barbara, CA, USA

First published on: 29 November 2010

To cite this Article Erb, Jack and Steigerwald, Douglas G.(2011) 'Accurately sized test statistics with misspecified conditional homoskedasticity', Journal of Statistical Computation and Simulation, 81: 6, 729 — 747, First published on: 29 November 2010 (iFirst)

To link to this Article: DOI: 10.1080/00949650903463574
URL: http://dx.doi.org/10.1080/00949650903463574

Journal of Statistical Computation and Simulation
Vol. 81, No. 6, June 2011, 729–747

Taylor & Francis
Taylor & Francis Group

Accurately sized test statistics with misspecified conditional homoskedasticity

Jack Erb* and Douglas G. Steigerwald

Department of Economics, University of California, Santa Barbara, CA 93106, USA

(*Received 1 April 2009; final version received 4 November 2009*)

We study the finite-sample performance of test statistics in linear regression models where the error dependence is of unknown form. With an unknown dependence structure, there is traditionally a trade-off between the maximum lag over which the correlation is estimated (the bandwidth) and the amount of heterogeneity in the process. When allowing for heterogeneity, through conditional heteroskedasticity, the correlation at far lags is generally omitted and the resultant inflation of the empirical size of test statistics has long been recognized. To allow for correlation at far lags, we study the test statistics constructed under the possibly misspecified assumption of conditional homoskedasticity. To improve the accuracy of the test statistics, we employ the second-order asymptotic refinement in Rothenberg [*Approximate power functions for some robust tests of regression coefficients*, Econometrica 56 (1988), pp. 997–1019] to determine the critical values. The simulation results of this paper suggest that when sample sizes are small, modelling the heterogeneity of a process is secondary to accounting for dependence. We find that a conditionally homoskedastic covariance matrix estimator (when used in conjunction with Rothenberg's second-order critical value adjustment) improves test size with only a minimal loss in test power, even when the data manifest significant amounts of heteroskedasticity. In some specifications, the size inflation was cut by nearly 40% over the traditional heteroskedasticity and autocorrelation consistent (HAC) test. Finally, we note that the proposed test statistics do not require that the researcher specify the bandwidth or the kernel.

Keywords: robust testing; test size; confidence interval estimation; heteroskedasticity; autocorrelation

2000 Mathematics Subject Classification: 62F03; 62F35; 62J05; 62M10; 91B84

1. Introduction

When forming test statistics for coefficients in linear regression models, it has become widely accepted to use the Newey–West covariance estimator to account for error serial correlation. The appeal of the Newey–West method [1] is that it allows for conditional heteroskedasticity, although at the cost of only admitting serial correlation at near lags. The inability to account for serial correlation at far lags leads to test statistics with empirical sizes that far exceed nominal sizes. To address this problem, Kiefer and Vogelsang [2] refine the asymptotic theory to more accurately model the admission of serial correlation at far lags. As Kiefer and Vogelsang show that the resultant non-Gaussian critical values increase with the admitted lag length, the desire to

*Corresponding author. Email: erb@econ.ucsb.edu

ISSN 0094-9655 print/ISSN 1563-5163 online
© 2011 Taylor & Francis
DOI: 10.1080/00949650903463574
http://www.informaworld.com

accommodate both conditional heteroskedasticity and correlation at far lags carries its own cost of considerably lengthening confidence intervals. In an effort to reduce this cost, we adjudge whether modelling the heteroskedasticity is secondary to accounting for dependence. To do so, we compare the performance of test statistics that allow for conditional heteroskedasticity with test statistics constructed under the (possibly) misspecified assumption of conditional homoskedasticity.

Driven by the desire to allow for general dependence in economic time series, White and Domowitz [3] develop a consistent standard error estimator under conditional heteroskedasticity. The key condition is that the maximum lag over which the serial correlation is estimated, the bandwidth, is an asymptotically negligible fraction of the sample size. Although the White–Domowitz estimator is consistent, it is not guaranteed to be positive definite. In response, Newey and West demonstrate that the introduction of a kernel that downweighs correlations as the lag length grows ensures that the consistent standard error estimator is also positive semi-definite.

With every solution, there comes another problem. While the Newey–West estimator is consistent and positive semi-definite, the estimated standard errors are often too small. As Andrews [4] demonstrates, if the errors exhibit substantive temporal dependence, then test statistics formed from the Newey–West standard error estimates have empirical size far in excess of nominal size (test statistics that reject too often). To mitigate size distortion, Andrews and Monahan [5] propose a two-step method, in which the first step consists of prewhitening the residuals by fitting a low-dimension process (such as a VAR(1)) to capture serial correlation at far lags. In the second step, the conditional heteroskedasticity is estimated at near lags.

Prewhitening the residuals prior to estimating conditional heteroskedasticity at near lags goes part way to resolving the problem of test over-rejection. In an effort to make further improvements, Kiefer and Vogelsang suggest forgoing the first-step prewhitening and estimating the conditional heteroskedasticity directly at both near and far lags. When including correlation at far lags, it is no longer tenable to assume that the bandwidth is an asymptotically negligible fraction of the sample size. In consequence, the (first-order) asymptotic distribution of resultant test statistics is not Gaussian. The alternative asymptotic distribution delivers simulated critical values that are considerably larger than their Gaussian counterparts. If only serial correlation at near lags is admitted, then the refined asymptotic critical values deliver size improvements in line with the improvements obtained by prewhitening. If serial correlation at all lags is admitted (note that the theory does not deliver an optimal bandwidth), then there are substantial further reductions in empirical size.

A key insight in previous research is the need to account for serial correlation at far lags to obtain more accurate coverage probabilities. Current methods to account for correlation at far lags are either completely general, as in Kiefer and Vogelsang, or specific, as in Andrews and Monahan.[1] While Kiefer and Vogelsang allow for conditional heteroskedasticity of unknown form, the cost is longer confidence intervals resulting in loss of power for associated test statistics. The low-dimension parametric method of Andrews and Monahan produces confidence intervals of more moderate length, but still suffers from high empirical size. It is therefore of interest to study methods that lie between the two, to adjudge the trade-off between size and power.

In contrast to Kiefer and Vogelsang, we propose broadening the first step of Andrews and Monahan by fitting a high-dimension process to capture serial correlation at all lags, while forgoing the second-step conditional heteroskedasticity estimation. The assumption underlying the method is that standard errors can be well approximated by a conditionally homoskedastic covariance matrix that is band diagonal. The band diagonals are not restricted to be related through a low-dimension process. To obtain size improvements, we too rely on asymptotic refinements, namely the second-order theory of Rothenberg [12]. The second-order theory yields critical values that adjust to incorporate the behaviour of the regressors.[2] As Rothenberg establishes, the bandwidth need not be an asymptotically negligible fraction of the sample size, so all correlation lags are included under conditional homoskedasticity. Further, the estimator is positive semi-definite by construction without need of a kernel.

We study the size and size-adjusted power of test statistics constructed under the three methods. We focus not only on hypothesis tests of a single parameter, but also on tests of multiple parameters to determine the impact of off-diagonal elements of the estimated covariance matrix. In Section 2, we present the quantities of interest and the models to be simulated. The conditional heteroskedasticity specifications allow us to investigate an additional observation in Rothenberg: namely, that the degree of correlation between the regressor under test and the conditional heteroskedasticity plays a key role in the empirical size. We examine the range of models typically used to assess the performance of confidence intervals under conditional heteroskedasticity. Results from the simulations are contained in Section 3. We provide an empirical application to the estimation techniques in Section 4.

2. Covariance estimators

To determine the finite-sample size and size-adjusted power of hypothesis tests constructed under the (potentially misspecified) assumption of conditional homoskedasticity, we employ a simulation model. Our simulation model is

$$Y_t = X_t'\beta + U_t \quad t = 1, \ldots, n, \tag{1}$$

where X_t and β are $k \times 1$ vectors. To allow for comparison with the findings in both Andrews as well as Andrews and Monahan, we set $k = 5$ where X_t contains a constant and four regressors. Conditional heteroskedasticity is introduced through a scale parameter that depends equally on each of the varying regressors

$$U_t = |X_t'\zeta| \times \tilde{U}_t,$$

where $\zeta = (0, 1/2, 1/2, 1/2, 1/2)'$ and $\{\tilde{U}_t\}$ is a sequence of possibly dependent random variables defined below. This specification of conditional heteroskedasticity is also employed by Andrews [4] and Andrews and Monahan [5] to demonstrate the superior performance of estimators that incorporate the conditional heteroskedasticity over the more traditional parametric covariance estimators.

The relative magnitude of conditional heteroskedasticity present in the model is controlled through the degree of serial correlation in the regressors and error. To capture serial correlation, the regressors and error are generated for each $t = 1, \ldots, n$ (and each $k = 2, \ldots, 5$) as

$$\begin{aligned} \tilde{U}_t &= \rho_U \tilde{U}_{t-1} + \varepsilon_t, \\ X_{kt} &= \rho_X X_{kt-1} + \eta_{kt}. \end{aligned}$$

The underlying errors, ε_t and η_{kt}, are mutually independent $N(0, 1)$ random variables.[3] The serial correlation parameters (ρ_U, ρ_X) take values in the set $\Lambda = \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$.[4]

Rothenberg derives the second-order asymptotic distribution of test statistics on coefficients of a linear regression under the assumption of conditional homoskedasticity. He finds (p. 1011) that as the correlation between the regressor under test and the scale parameter increases, the inflation of the empirical size of the test statistics increases. Therefore, we employ a second conditional heteroskedasticity specification in which the scale parameter depends on only one of the varying regressors. To isolate the impact noted by Rothenberg, we consider two values for the parameter ζ, specifically $\zeta = (0, 1, 0, 0, 0)'$ and $\zeta = (0, 0, 1, 0, 0)'$, while testing the hypothesis that the first varying regressor is equal to zero. The first specification provides the highest level of correlation between the regressor under test and the error scale, while in the second specification the correlation between the regressor and scale is zero.[5,6]

We focus not only on hypothesis tests of a single coefficient, but also on tests of multiple coefficients. Our hypothesis tests of multiple coefficients are designed to assess the effect of including off-diagonal elements of the covariance matrix. To capture this effect, we consider the single restriction imposed by the hypothesis $H_0 : \beta_2 - \beta_3 = 0$. In the multiple coefficient tests, we consider values of the heteroskedasticity parameter, ζ, equal to $(0, 1/2, 1/2, 1/2, 1/2)'$, $(0, 1, 0, 0, 0)'$, and $(0, 0, 0, 1, 0)'$.

To construct test statistics for hypotheses concerning β, an estimator of the (conditional) variance of the ordinary least-squares (OLS) estimator, B, is needed. The variance of the (ordinary) least-squares estimator, conditional on $X = (X_1, \ldots, X_n)'$, is

$$\begin{aligned}&\mathrm{Var}(n^{1/2}(B-\beta)|X)\\ &= \Big(n^{-1}\sum_{t=1}^{n} X_t X_t'\Big)^{-1} n^{-1}\sum_{s=1}^{n}\sum_{t=1}^{n} E(U_s X_s U_t X_t' \,|\, X)\Big(n^{-1}\sum_{t=1}^{n} X_t X_t'\Big)^{-1}.\end{aligned}$$

The key component for estimation is $J = n^{-1}\sum_{s=1}^{n}\sum_{t=1}^{n} E(U_s X_s U_t X_t' | X)$.

We consider three estimators of J.[7] The first is the heteroskedasticity–autocorrelation-consistent estimator introduced by Newey and West

$$\hat{J}_{\mathrm{hac}} = \frac{1}{n-5}\left[\sum_{t=1}^{n} \hat{U}_t^2 X_t X_t' + \sum_{j=1}^{m}\left(1 - \frac{j}{m+1}\right)\sum_{t=j+1}^{n} \hat{U}_t \hat{U}_{t-j}(X_t X_{t-j}' + X_{t-j} X_t')\right],$$

where $\{\hat{U}_t\}_{t=1}^{n}$ is the OLS residual vector.[8] The value of m determines the maximum lag length at which the conditional heteroskedasticity is estimated. As m controls the number of far lags that enter the estimator, sample-based selection of m is very important in controlling the size of test statistics. The value of the bandwidth is allowed to vary across simulations and is chosen according to the automatic selection procedure developed by Andrews.

We also focus on a variant of the heteroskedasticity–autocorrelation-consistent estimator, discussed by Andrews and Monahan. To reduce the bias in $\hat{J}_{\mathrm{hac}}$, Andrews and Monahan advocate separating the variance estimation into three steps. First, estimate the temporal correlation (at far lags) by fitting a vector autoregression to $\{\hat{V}_t\}$ (where $\hat{V}_t = \hat{U}_t X_t$), which yields the prewhitened residuals $\{\tilde{V}_t\}$. (In our implementation, we fit a vector autoregression of order 1.) Second, construct the variance estimator with the prewhitened residuals

$$\tilde{J}_{\mathrm{pw}} = \frac{1}{n-5}\left[\sum_{t=1}^{n} \tilde{V}_t \tilde{V}_t' + \sum_{j=1}^{m}\left(1 - \frac{j}{m+1}\right)\sum_{t=j+1}^{n} (\tilde{V}_t \tilde{V}_{t-j}' + \tilde{V}_{t-j} \tilde{V}_t')\right].$$

The last step is to recolour the estimator $\tilde{J}_{\mathrm{pw}}$ to obtain the prewhitened variance estimator, $\hat{J}_{\mathrm{pw}}$, according to

$$\hat{J}_{\mathrm{pw}} = \hat{C}\tilde{J}_{\mathrm{pw}}\hat{C}', \quad \text{where } \hat{C} = \left(I - \sum_{s=1}^{p} \hat{A}_s\right)^{-1}.$$

Here, $\{\hat{A}_s\}_{s=1}^{p}$ are the estimated coefficient matrices from a pth-order vector autoregression of V_t.[9]

With the reduction in correlation brought about by the first-step prewhitening, there is less need to select a large value of m.[10] Although the variance of $\hat{J}_{\mathrm{pw}}$ exceeds the variance of $\hat{J}_{\mathrm{hac}}$, Andrews and Monahan find that use of the prewhitened residuals reduces the (downward) bias of $\hat{J}_{\mathrm{hac}}$. The downward bias in the estimated standard errors is reduced to such an extent that, despite a loss of precision in estimating the standard errors, the coverage probabilities of confidence intervals are increased.

Constructing accurately sized tests with each of the above estimators of J remains a problem. However, the parametric prewhitening in $\hat{J}_{\mathrm{pw}}$ does improve the size of test statistics relative to $\hat{J}_{\mathrm{hac}}$. Therefore, it may be the case that increasing the richness of the first step, in which temporal correlation is accounted for, lessens the importance of the second step, in which conditional heteroskedasticity is accounted for. Rather than assume a low-dimension parametric model for temporal correlation – such as a small-order, autoregressive model – one could assume that the errors are generated by a conditionally homoskedastic (stationary stochastic) process with nothing further known about the autocorrelation function. Under this assumption,

$$n^{-1}\sum_{s=1}^{n}\sum_{t=1}^{n}E(U_sX_sU_tX_t'|X)=n^{-1}\sum_{s=1}^{n}\sum_{t=1}^{n}\delta_{|t-s|}X_sX_t',$$

where $\delta_{|t-s|}=E(U_sU_t|X)$ depends only on $|t-s|$. The third estimator of J, which is consistent if the errors are conditionally homoskedastic, is

$$\hat{J}_{\mathrm{cho}}=\frac{1}{n-5}\sum_{s=1}^{n}\sum_{t=1}^{n}\hat{\delta}_{|t-s|}X_sX_t',$$

where (for $t>s$) $\hat{\delta}_{|t-s|}=\frac{1}{n}\sum_{t=s+1}^{n}\hat{U}_t\hat{U}_{t-s}$. (If $s>t$, simply switch the values of t and s in the formula.)[11]

Under traditional asymptotics, all of the estimators of J lead to Gaussian limit distributions in testing situations. However, extensions to the asymptotic theory are available for two of the estimators. For $\hat{J}_{\mathrm{hac}}$, Kiefer and Vogelsang have developed an alternative limit theory based on the assumption that the fraction of lags that appear in the estimator, m/n, is not asymptotically negligible.[12] The critical values that arise from the alternative limit theory can be considerably larger than the standard Gaussian critical values. As these critical values depend on m, we report simulation results for two sets of values of the bandwidth. The first uses the Andrews automatic bandwidth procedure to compute the bandwidth and the second sets the bandwidth equal to the sample size ($m=n$).[13,14]

For $\hat{J}_{\mathrm{cho}}$, Rothenberg provides critical values based on a higher-order asymptotic refinement under strict exogeneity. If cv denotes the critical value from the first-order Gaussian approximation, then Rothenberg's second-order theory delivers the adjusted critical value

$$\mathrm{cv}^{\mathrm{R}}=\mathrm{cv}\left(1+\frac{1}{n}f(X,\hat{U})\right).^{15}$$

His asymptotic refinements indicate that the Gaussian critical values should generally be increased (as f is generally greater than zero), although the precise form of his covariance estimator differs slightly from $\hat{J}_{\mathrm{cho}}$.[16] Because the adjusted critical value is a function of $(X,\hat{U})$, the adjusted critical value is correlated with the estimated standard error. If this correlation is negative, then the critical value adjusts to the magnitude of the standard error and lessens the length of estimated confidence intervals. Such an adjustment feature can lead to a test statistic with large gains in size at the cost of only small reductions in size-adjusted power.

3. Simulation results

For the simulations, we construct Y_t according to Equation (1) with $\beta=\mathbf{0}$. Letting c be a 5×1 vector of constants that selects the parameters under test, we construct the test statistic for the hypothesis $H_0 : c'\beta = 0$ according to

$$t = [c'V_B c]^{-1/2} \cdot \sqrt{n} c' B,$$

where B is the OLS estimate of β and V_B is the 5×5 sample analogue of $\mathrm{Var}(n^{1/2}(B - \beta)|X)$. ($V_B$ is constructed for each of the variance estimators in Section 2.) The selected critical values are for a two-sided test with 5% nominal size. The sample size is $n = 50$ in all models to allow for direct comparison with the simulation results presented in recent papers, such as Kiefer and Vogelsang [2] and Phillips *et al.* [14]. Each experiment consists of 50,000 replications. While there are a variety of statistics that can be used to assess the finite-sample performance of the variance estimators, we follow the convention of more recent authors and focus our attention on the finite-sample size and size-adjusted power of test statistics.

It is well known that with any hypothesis test, there is a trade-off between size and power. Because each test employs the OLS estimator as a point estimate, improvements in size will generally be accompanied by a corresponding decrease in power. However, each test varies in either the variance estimate, critical value, or both. Consequently, there is the possibility that a particular estimator may display more accuracy in terms of both improved test size and higher power against alternatives.

3.1. *Single parameter tests*

We first study $H_0 : c'\beta = 0$, with $c = (0, 1, 0, 0, 0)'$, and so test whether the coefficient on the first non-constant regressor is significantly different from zero. This choice of c leads to a standard error estimate generated from only one of the diagonal elements of V_B. In the tables that follow, the test statistics are referenced by the covariance matrix estimator the test employs. Thus, hac denotes the t-statistic when V_B is constructed according to $\hat{J}_{\text{hac}}$ and evaluated with standard asymptotic critical values. Critical values from the asymptotic refinements are indicated by superscripts, so cho$^{\text{R}}$ is the t-statistic constructed with $\hat{J}_{\text{cho}}$ and evaluated with Rothenberg's second-order critical values. In a similar fashion, hac$^{\text{KV}}$ indicates the use of $\hat{J}_{\text{hac}}$ with the Kiefer–Vogelsang asymptotic approximation to generate test critical values and a bandwidth determined by the Andrews automatic selection procedure. As the Kiefer-Vogelsang approximation allows the bandwidth to equal the sample size, we denote this statistic as hac$^{\text{KV}n}$.

Table 1 reports the finite-sample empirical size of each test statistic when the AR(1) errors are overlaid with multiplicative heteroskedasticity entering from all four non-constant regressors, i.e. $\zeta = (0, 1/2, 1/2, 1/2, 1/2)'$. The nominal size for each test is 0.05. As with previous authors, we find that the traditional hac test performs quite poorly in terms of test size, especially when the dependence in the data is strong. Indeed, when $\rho_X = \rho_U = 0.9$, the hac test rejects the null hypothesis 38% of the time. This is quite unsettling considering that the applied researcher will be making inference based on a nominal size of 5%.

As the test statistics all employ the same point estimate in the numerator, improvements in test size will be achieved by either increasing the standard error in the denominator or widening the test critical values. In column 2, we see that prewhitening residuals reduces test size by inflating the estimated standard errors, although over-rejection of the null remains a problem. In the $\rho_X = \rho_U = 0.9$ case, the empirical size drops to 0.32, a 16% size gain. However, attempting to remove correlation at far lags by prewhitening when serial correlation in the data is weak can inflate test size even more than using the HAC estimator, which can be seen by comparing columns 1 and 2 when either ρ_X or ρ_U is less than 0.3.

Column 3 shows that use of the Kiefer–Vogelsang asymptotic refinements reduces size inflation by widening the test critical values. The new limit theory leads to critical values that typically take on (absolute) values in the range of 2.0–4.81, depending on the chosen bandwidth for the

Table 1. Empirical size: heteroskedastic AR(1) errors, $\zeta = (0, 1/2, 1/2, 1/2, 1/2)$.

ρ_X	ρ_U	(1), hac	(2), pw	(3), hac$^{\mathrm{KV}}$	(4), hac$^{\mathrm{KV}n}$	(5), cho	(6), cho$^{\mathrm{R}}$
0.9	0.9	0.3808	0.3235	0.3171	0.2332	0.3192	0.2367
	0.7	0.2834	0.2351	0.2327	0.1686	0.2371	0.1641
	0.5	0.2126	0.1803	0.1766	0.1322	0.1941	0.1306
	0.3	0.1655	0.1501	0.1391	0.1102	0.1668	0.1076
	0.1	0.1310	0.1308	0.1088	0.0934	0.1479	0.0906
	0.0	0.1213	0.1254	0.0987	0.0892	0.1403	0.0809
0.7	0.9	0.2859	0.2460	0.2393	0.1680	0.2481	0.1854
	0.7	0.2352	0.2025	0.1970	0.1414	0.2087	0.1508
	0.5	0.1933	0.1726	0.1645	0.1217	0.1844	0.1326
	0.3	0.1554	0.1469	0.1324	0.1034	0.1591	0.1136
	0.1	0.1327	0.1332	0.1108	0.0903	0.1440	0.1018
	0.0	0.1209	0.1260	0.1007	0.0886	0.1345	0.0929
0.5	0.9	0.2108	0.1880	0.1745	0.1208	0.1948	0.1493
	0.7	0.1836	0.1669	0.1557	0.1080	0.1711	0.1295
	0.5	0.1621	0.1519	0.1385	0.1019	0.1569	0.1204
	0.3	0.1407	0.1382	0.1214	0.0927	0.1434	0.1077
	0.1	0.1226	0.1275	0.1054	0.0831	0.1342	0.1008
	0.0	0.1159	0.1244	0.0994	0.0819	0.1285	0.0963
0.3	0.9	0.1592	0.1492	0.1355	0.0949	0.1579	0.1273
	0.7	0.1473	0.1419	0.1261	0.0924	0.1471	0.1153
	0.5	0.1367	0.1353	0.1194	0.0871	0.1409	0.1100
	0.3	0.1239	0.1277	0.1080	0.0819	0.1281	0.0994
	0.1	0.1167	0.1232	0.1024	0.0793	0.1282	0.0996
	0.0	0.1142	0.1240	0.0996	0.0783	0.1237	0.0967
0.1	0.9	0.1215	0.1244	0.1046	0.0749	0.1295	0.1073
	0.7	0.1187	0.1236	0.1024	0.0777	0.1285	0.1036
	0.5	0.1199	0.1277	0.1047	0.0796	0.1289	0.1031
	0.3	0.1138	0.1227	0.0995	0.0758	0.1239	0.0997
	0.1	0.1091	0.1180	0.0957	0.0745	0.1189	0.0937
	0.0	0.1094	0.1186	0.0954	0.0736	0.1195	0.0945
0.0	0.9	0.1105	0.1190	0.0948	0.0719	0.1225	0.1011
	0.7	0.1087	0.1181	0.0941	0.0728	0.1225	0.0995
	0.5	0.1089	0.1191	0.0949	0.0745	0.1203	0.0958
	0.3	0.1069	0.1178	0.0945	0.0747	0.1181	0.0949
	0.1	0.1078	0.1196	0.0947	0.0759	0.1200	0.0947
	0.0	0.1059	0.1172	0.0925	0.0729	0.1190	0.0939

HAC estimator. When the bandwidth is chosen according to the Andrews procedure, the size gains offered by the Kiefer–Vogelsang asymptotics are slightly larger than those achieved by prewhitening, but are typically comparable when the serial correlation is strong.

Similar to prewhitening, we find that $\hat{J}_{\mathrm{cho}}$ delivers larger standard errors that improve the finite-sample size of the resulting test statistics, and similar to the Kiefer–Vogelsang asymptotics, we find that the second-order critical value refinement of Rothenberg further improves the test size by increasing the critical values. Columns 5 and 6 of Table 1 present the empirical size of test statistics when $\hat{J}_{\mathrm{cho}}$ is used in conjunction with either the standard normal critical values or the second-order critical values of Rothenberg, respectively. Even under misspecification, the cho$^{\mathrm{R}}$ test delivers more substantial size reductions than both the pw and hac$^{\mathrm{KV}}$ tests. For $\rho_X = \rho_U = 0.9$, the empirical size of the cho$^{\mathrm{R}}$ test is 0.24. Although the actual size remains significantly larger than the 5% nominal level, the cho$^{\mathrm{R}}$ test reduces size inflation by more than one-third of the level of the hac test. Moreover, the improved accuracy of the cho$^{\mathrm{R}}$ test over the more conventional HAC tests continues to hold for very low levels of serial correlation. For $\rho_X = \rho_U = 0.1$, the second-order adjusted, homoskedastic estimator improves test size accuracy relative to the Newey–West estimator by about 14% or a drop in raw size from 0.1091 to 0.0937.[17]

Recall that for a fixed value of ρ_X, the heteroskedastic component of the error becomes more pronounced as ρ_U decreases. That the conditionally homoskedastic estimator retains an advantage in test size in this case is quite remarkable. The favourable performance of cho$^{\mathrm{R}}$ is primarily due to the adaptability of the critical value to the data-generating process. The second-order theory for $\hat{J}_{\mathrm{cho}}$ delivers critical values that, while typically larger than their Gaussian counterparts, adjust with the regressors and residuals in such a way that the critical value increases when the estimated standard error is small and decreases when the estimated standard error is large. This negative correlation between the standard error and critical value serves as a hedge in cases where over-rejections of the null are most likely to occur. As can be seen from column 5 of Table 1, $\hat{J}_{\mathrm{cho}}$ is indeed influenced by the heteroskedasticity as the cho test performs more poorly than the hac test when ρ_U drops below 0.3. However, when the $\hat{J}_{\mathrm{cho}}$ estimator is used in conjunction with the second-order critical values, the cho$^{\mathrm{R}}$ test retains a size advantage over the hac and pw tests for all values of ρ_X and ρ_U, and a size advantage over hac$^{\mathrm{KV}}$ for all but the smallest values of ρ_X and ρ_U.[18]

Table 2 reports the mean and variance of Rothenberg's second-order critical value, as well as its correlation with the estimated standard error, when ρ_X is fixed at 0.9. While the mean of the critical value remains relatively constant as ρ_U decreases, the variation increases and the negative correlation with the estimated standard error becomes more pronounced. The adaptability of the critical value allows the conditionally homoskedastic estimator to maintain size improvements over the hac, hac$^{\mathrm{KV}}$, and pw tests, even as the heteroskedasticity in the process becomes more pronounced.

It is important to note that the asymptotic theory put forth by Kiefer and Vogelsang does not restrict the bandwidth to be small relative to the sample size in order for the testing procedure to be valid. When the bandwidth is set equal to the sample size, the considerable downward bias of the Newey–West estimator is offset by an adjusted critical value of 4.81, which is almost two and a half times the critical value of the standard normal distribution. In comparison, the average critical value for the hac$^{\mathrm{KV}}$ test (as selected by the Andrews method) ranges from 2.04 when the temporal dependence is low to 2.26 when the dependence is high, and the average critical value for the cho$^{\mathrm{R}}$ test ranges from 2.11 when the dependence is low to 2.37 when the dependence is high. Column 4 in Table 1 shows that for more moderate levels of dependence, further reductions in test size are achieved with the Kiefer–Vogelsang critical values if the bandwidth is fixed and equal to the sample size, though the improvements are small. However, such drastic inflation of the critical value is sure to decrease the probability of rejecting the hypothesis for all values of β, and the slight size improvements of the hac$^{\mathrm{KV}n}$ test over the cho$^{\mathrm{R}}$ test prove to be extremely costly in terms of test power.

In order to ascertain the power of the various testing procedures, we again simulate (1) and compute the size-adjusted power against alternative values of β. It is important to first note that size adjustment is not possible in practice as the underlying data-generating process is typically not known to the researcher. However, the technique may be useful in developing methods as it is useful in comparing the power of tests that do not have the same finite-sample size. In our simulations, the stochastic process is known, and a size-adjusted critical value for each specification can be calculated via simulation.

Table 2. Performance of Rothenberg's second-order-adjusted critical value under heteroskedasticity.

ρ_X	ρ_U	(1), Mean	(2), Variance	(3), Corr. with $\sqrt{c'V_Bc}$
0.9	0.9	2.37	0.04	−0.15
	0.7	2.34	0.04	−0.22
	0.5	2.32	0.05	−0.27
	0.3	2.32	0.05	−0.32
	0.1	2.33	0.08	−0.34
	0.0	2.34	0.08	−0.37

Figure 1 plots the upper half of the size-adjusted power functions for the test statistics under the AR(1) specification presented in Table 1 when $\rho_X = \rho_U$. We compute size-adjusted power as the fraction of test rejections that arise when the true value of β_2 is different from the hypothesized null value of zero. Specifically, we set $\beta = \psi \times (0, 1, 0, 0, 0)'$ where ψ is chosen as a set of 11, equally spaced points from zero to some upper bound for which the estimated power for all tests is roughly one.[19] We simulate the power of test statistics using 10,000 replications for each value of ψ. The size-adjusted critical values are also computed via simulation methods using 50,000 replications. As each estimator gives rise to only one finite-sample distribution, the size-adjusted critical values and size-adjusted power curves for the hac and hac$^{\mathrm{KV}}$ tests are equivalent (as are those for the cho and cho$^{\mathrm{R}}$ tests).

Returning to our discussion of the hac$^{\mathrm{KV}n}$ test, the most striking feature of Figure 1 is the extent to which the power of the hac$^{\mathrm{KV}n}$ test lags the power of the other tests. As an example, consider the case where $\beta_2 = 2.4$ and the serial correlation parameters are equal to 0.9. The size-adjusted power of the hac test is 0.72, while the power of the cho$^{\mathrm{R}}$ test is 0.70. Recall the empirical sizes of the two tests were 0.38 and 0.24, respectively, indicating a 37% reduction in test size is achieved with approximately a 3% decrease in test power. As noted above, the empirical size is further improved to 0.23 when using the hac$^{\mathrm{KV}n}$ test. However, the power of the test falls to 0.59, and we see that the slight additional improvement in the size of the hac$^{\mathrm{KV}n}$ test comes at the cost of reducing the power by 16%. Last, the figure shows that the cho$^{\mathrm{R}}$ test provides size-adjusted power that is quite similar to both the hac and pw tests (with minimal crossing).

In general, the power functions for other values of ρ_X and ρ_U are similar in shape and relative performance to those presented in Figure 1.[20] The HAC, prewhitened, and conditionally

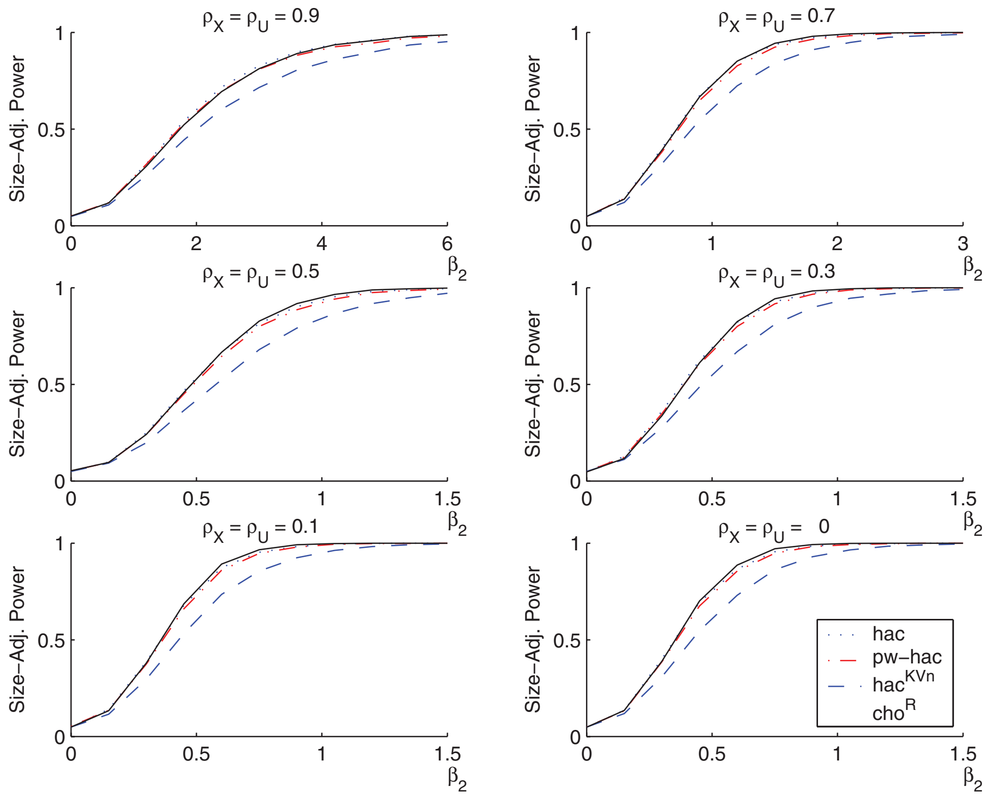

Figure 1. Size-adjusted power: $\zeta = (0, 1/2, 1/2, 1/2, 1/2)'$ $(H_0 : \beta_2 = 0)$.

homoskedastic estimators give rise to tests with very similar power, while the hac$^{\mathrm{KV}n}$ test exhibits power that is substantially lower across the range of alternatives.

An unfortunate feature of the hac and pw tests is that rejection depends heavily upon user-choice parameters (including bandwidth, weighting kernel, and prewhitening model). For this reason, Andrews and others have developed data-driven, 'optimal' bandwidth selection procedures in an attempt to make the estimation process more automatic.[21] The Kiefer–Vogelsang asymptotics eliminate the need of choosing an 'optimal' bandwidth by allowing the critical values to adjust with the chosen bandwidth. In this way, the choice of bandwidth is incorporated directly into the testing problem. In practice, however, the researcher must still chose a bandwidth, and different choices may give rise to very different inferences on the parameter under test.[22]

To illustrate the problem that bandwidth selection could cause, Table 3 shows the probability that the hac$^{\mathrm{KV}}$ test rejects the null hypothesis for at least one bandwidth in a given replication. That is, for any given replication, we constructed a test statistic and critical value pair for each value of the bandwidth between $m = 5$ and $m = n = 50$.[23] If the null hypothesis was rejected for one or more of the test statistic/critical value pairs, the entire trial was considered as if it rejected the null hypothesis. We then repeated the process 50,000 times and found the fraction of replications which produced at least one rejection. The probability of finding at least one bandwidth value that produces a test rejection in the hac$^{\mathrm{KV}}$ test (Table 3) is much larger than the empirical size of the hac$^{\mathrm{KV}}$ tests in column 3 of Table 1. In fact, the probabilities are often larger in magnitude than the sizes of the traditional hac test in column 1 of Table 1. Clearly, the failure to properly account for the pre-test estimation of the bandwidth results in further size inflation and negates the advantages of the hac$^{\mathrm{KV}}$ test over the more traditional tests.

The problem of nuisance parameters in the testing process highlights an important advantage of $\hat{J}_{\mathrm{cho}}$ over other estimators. First, $\hat{J}_{\mathrm{cho}}$ estimates the correlation at all lags, eliminating the necessity of choosing a particular bandwidth. Second, there is no weighting kernel or prewhitening filter involved in the estimation. And third, the second-order critical value adjustment is 'automatic' in the sense that it is data dependent and can be implemented without any choices by the user.

While the cho$^{\mathrm{R}}$ test performs favourably under the specification in Table 1 when compared with more traditional tests commonly used in the literature, it is also important to evaluate the test under varying degrees of serial correlation and heteroskedasticity.[24] One way to alter the degree of heteroskedasticity in the data is to adjust the impact of individual regressors on the error scale. Recall that the error term is generated according to

$$U_t = |X_t'\zeta| \times \tilde{U}_t.$$

Rather than set $\zeta = (0, 1/2, 1/2, 1/2, 1/2)'$, as in the previous specifications, we now set $\zeta = (0, 1, 0, 0, 0)'$ and allow the heteroskedasticity to arise from scaling the homoskedastic error

Table 3. Probability the hac$^{\mathrm{KV}}$ test rejects the null hypothesis for at least one value of the bandwidth.

	AR(1) regressors and errors – $\zeta = (0, 1/2, 1/2, 1/2, 1/2)$					
	ρ_X					
ρ_U	0.9	0.7	0.5	0.3	0.1	0.0
0.9	0.3662	0.2788	0.2164	0.1724	0.1442	0.1336
0.7	0.2776	0.2310	0.2020	0.1654	0.1418	0.1300
0.5	0.2258	0.2030	0.1756	0.1482	0.1454	0.1256
0.3	0.1916	0.1752	0.1620	0.1448	0.1342	0.1174
0.1	0.1766	0.1542	0.1478	0.1344	0.1380	0.1284
0.0	0.1468	0.1458	0.1362	0.1250	0.1328	0.1412

term, $\tilde{U}_t$, by the absolute value of the first non-constant regressor. As we are testing the hypothesis $H_0 : \beta_2 = 0$, it may matter whether the heteroskedasticity is brought about by the first non-constant regressor or some other regressor in the model. For this reason, we also report the empirical size when the error is scaled by the absolute value of the second non-constant regressor, $\zeta = (0, 0, 1, 0, 0)'$.

The upper half of Table 4 shows the empirical size of tests when the heterogeneity arises from the first non-constant regressor where, for brevity, we report only the case where $\rho_U = \rho_X$. While the cho$^{\text{R}}$ test performs comparably to other robust tests when the amount of serial correlation is high, it no longer exhibits a size advantage, and it becomes progressively disadvantaged as the values of ρ_U and ρ_X fall towards zero. However, when testing the coefficient on the second non-constant regressor, the results change dramatically. The lower half reports the empirical size of the t-tests when $\zeta = (0, 0, 1, 0, 0)'$ and the heteroskedasticity arises from a regressor other than the regressor under test. In this case, the cho$^{\text{R}}$ test offers a considerable size advantage over all the other tests, even as the values of ρ_X and ρ_U approach zero. In fact, as the serial correlation parameters fall to zero, the size of the cho$^{\text{R}}$ test falls below the 5% nominal level.

This table would appear to confirm the observation in Rothenberg that the degree of correlation between the regressor under test and the error variance has a substantial impact on inflating the empirical size of test statistics. While this is true for all tests under examination, the size distortion is especially pronounced for cho$^{\text{R}}$. In practice, the form in which the heteroskedasticity enters the model appears to be of considerable importance when performing tests of hypotheses.

Figures 2 and 3 plot the corresponding size-adjusted power curves for the AR(1) process with $\rho_X = \rho_U$ when $\zeta = (0, 1, 0, 0, 0)'$ and $\zeta = (0, 0, 1, 0, 0)'$, respectively. Once again, we see cho$^{\text{R}}$ performs comparably to hac and hac$^{\text{KV}}$ in terms of size-adjusted power, and all three outperform hac$^{\text{KV}n}$ by a considerable margin.

3.2. *Multiple parameter tests*

We now briefly turn our attention to testing hypotheses that involve multiple parameters. These multi-parameter tests incorporate at least one off-diagonal element of V_B in computing the standard error. Specifically, we examine the performance of tests under the null hypothesis $H_0 : \beta_2 - \beta_3 = 0$ (or $H_0 : c'\beta = 0$ for $c = (0, 1, -1, 0, 0)'$) for each of the heteroskedastic specifications listed above. The results are found in Table 5, where once again we report only the results from setting $\rho_U = \rho_X$.

Table 4. Empirical size – heteroskedastic AR(1) errors.

$\rho_X = \rho_U$	(1), hac	(2), pw	(3), hac$^{\text{KV}}$	(4), hac$^{\text{KV}n}$	(5), cho	(6), cho$^{\text{R}}$
$\zeta = (0, 1, 0, 0, 0)'$						
0.9	0.4452	0.3959	0.3820	0.2781	0.4851	0.4023
0.7	0.2672	0.2313	0.2287	0.1625	0.3748	0.3090
0.5	0.1834	0.1708	0.1609	0.1142	0.3169	0.2674
0.3	0.1417	0.1440	0.1253	0.0935	0.2791	0.2399
0.1	0.1271	0.1369	0.1123	0.0839	0.2697	0.2321
0.0	0.1244	0.1335	0.1108	0.0829	0.2945	0.2312
$\zeta = (0, 0, 1, 0, 0)'$						
0.9	0.3191	0.2561	0.2566	0.1883	0.2136	0.1402
0.7	0.1925	0.1632	0.1580	0.1120	0.1168	0.0756
0.5	0.1319	0.1234	0.1104	0.0822	0.0824	0.0559
0.3	0.0994	0.1049	0.0844	0.0672	0.0670	0.0473
0.1	0.0886	0.0988	0.0761	0.0610	0.0606	0.0439
0.0	0.0866	0.0978	0.0748	0.0624	0.0607	0.0439

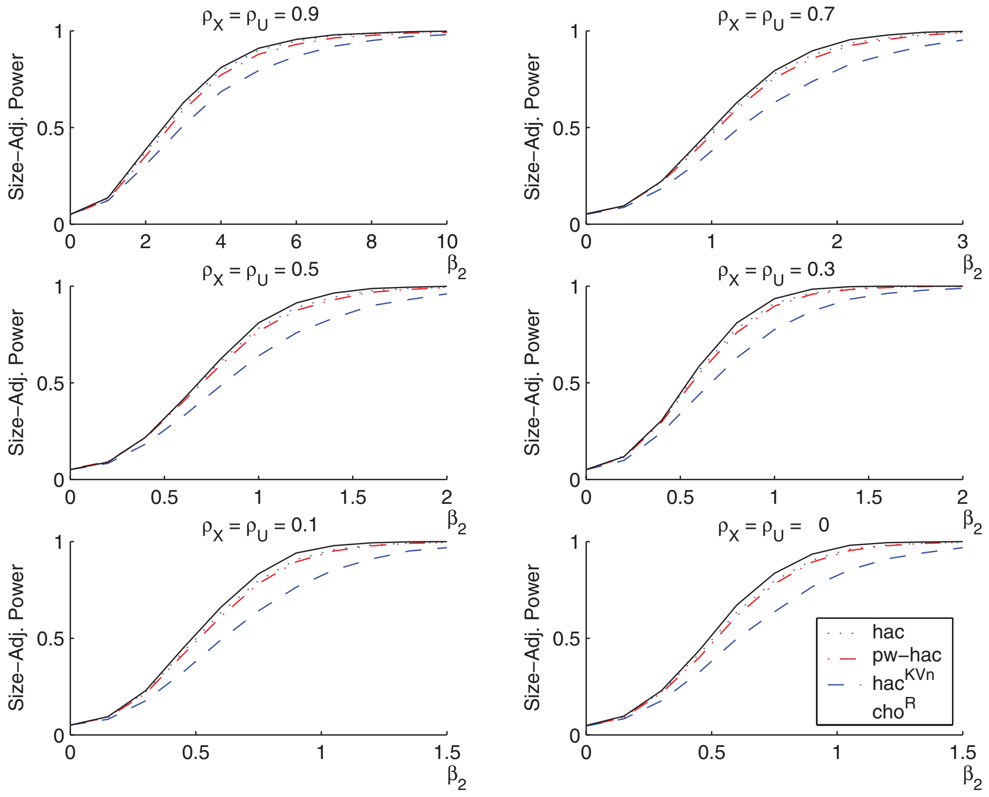

Figure 2. Size-adjusted power: $\zeta = (0, 1, 0, 0, 0)'$ ($H_0 : \beta_2 = 0$).

In testing the hypothesis $\beta_2 - \beta_3 = 0$, the cho^R test provides the best empirical size under all heteroskedasticity specifications and under all degrees of dependence, excepting the case where $\zeta = (0, 1, 0, 0, 0)'$. The size advantage of the cho^R tests in the first and third heteroskedasticity specifications is quite large. The empirical size of the cho^R test is cut by more than half the level of the traditional estimators, and it even exhibits better size than the hac^{KVn} test. In both cases, the empirical size approaches the nominal size rather quickly as the serial correlation decreases. When the heteroskedasticity enters solely from the regressors under test, the cho^R test is not as disadvantaged as in previous specifications, but still cannot be given a strong recommendation.

Last, note that in the model under study, the testing problem is made easier by the introduction of the second parameter to the hypothesis. Comparing the relevant rows of Table 5 with the corresponding rows in previous tables where $\rho_U = \rho_X$, we see that the empirical sizes of these tests are lower than the sizes of the single parameter tests across the full range of estimators. The same is also true for values of $\rho_U \neq \rho_X$ (though the results are not reported here). The size-adjusted power functions for these multi-parameter tests follow the same patterns as the power functions of previous tests and can be found in Erb and Steigerwald [13].

4. Empirical application

To understand the impact of our proposed standard error estimator on applied research, we revisit Lustig and Verdelhan [17] who use the Newey–West standard error estimator in their study of

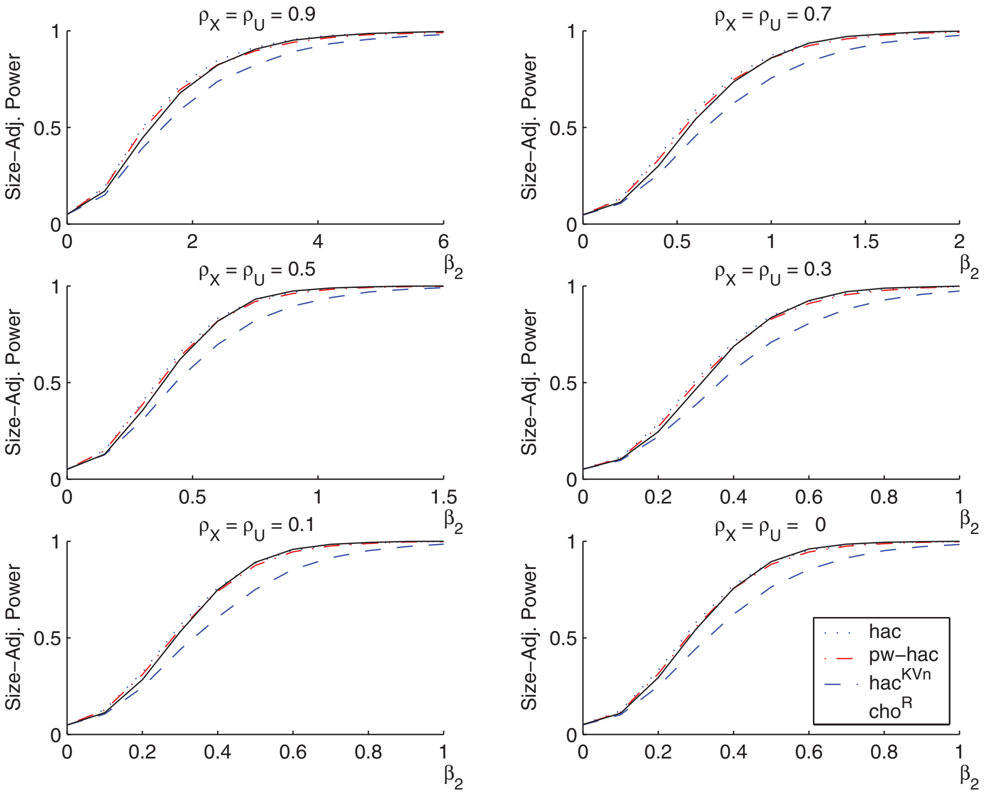

Figure 3. Size-adjusted power: $\zeta = (0, 0, 1, 0, 0)'$ ($H_0 : \beta_2 = 0$).

Table 5. Empirical size of multi-parameter tests – het. AR(1) errors

	$H_0: \beta_2 - \beta_3 = 0$					
$\rho_X = \rho_U$	(1), hac	(2), pw	(3), hac^{KV}	(4), hac^{KVn}	(5), cho	(6), cho^{R}
$\zeta = (0, 1/2, 1/2, 1/2, 1/2)'$						
0.9	0.3220	0.2589	0.2560	0.1858	0.2118	0.1389
0.7	0.1930	0.1628	0.1561	0.1111	0.1157	0.0723
0.5	0.1324	0.1252	0.1109	0.0837	0.0835	0.0576
0.3	0.1016	0.1066	0.0859	0.0675	0.0668	0.0474
0.1	0.0902	0.0993	0.0771	0.0632	0.0621	0.0461
0.0	0.0840	0.0950	0.0712	0.0583	0.0593	0.0425
$\zeta = (0, 1, 0, 0, 0)'$						
0.9	0.4185	0.3653	0.3544	0.2600	0.3951	0.3091
0.7	0.2513	0.2184	0.2142	0.1514	0.2768	0.2142
0.5	0.1723	0.1620	0.1505	0.1073	0.2185	0.1739
0.3	0.1346	0.1372	0.1168	0.0882	0.1899	0.1549
0.1	0.1213	0.1307	0.1064	0.0803	0.1752	0.1465
0.0	0.1181	0.1293	0.1043	0.0793	0.1772	0.1458
$\zeta = (0, 0, 0, 1, 0)'$						
0.9	0.3209	0.2545	0.2540	0.1862	0.2133	0.1388
0.7	0.1883	0.1589	0.1519	0.1098	0.1163	0.0753
0.5	0.1356	0.1258	0.1134	0.0812	0.0837	0.0572
0.3	0.1026	0.1073	0.0883	0.0721	0.0678	0.0482
0.1	0.0885	0.0976	0.0751	0.0643	0.0624	0.0446
0.0	0.0887	0.0986	0.0765	0.0636	0.0635	0.0463

foreign currency risk premia. Lustig and Verdelhan propose that the excess returns of foreign currency markets can be partially explained as a premium on aggregate consumption growth risk in US markets. The premium arises because returns from high-interest-rate countries tend to be low when (US aggregate) consumption growth is low. Conversely, the assets of the low-interest-rate countries negatively covary with consumption growth and so can serve as a hedge against US aggregate consumption growth risk.

To verify their proposition, Lustig and Verdelhan study a panel of countries from which they first assign each country to one of eight currency portfolios. The portfolios, which are formed by sorting on the short-term risk-free interest rate of the country, are rebalanced every period so that the countries with the lowest interest rates are in the first portfolio and countries with the highest interest rates are in the eighth portfolio. With annual data from 1953–2002, they estimate the covariances, which underpin their analysis, between excess currency returns and the growth rate of consumption. For each of the portfolios, estimates are obtained from regressions of the form

$$R_{t+1} = \alpha + \beta f_t + U_t,$$

where R_{t+1} is the annual excess return on the portfolio and f_t the growth of consumption in the previous year. Separate regressions are run for durable and non-durable consumption.

The first row of Table 6 replicates a subset of the results found in Table 6 of Lustig and Verdelhan, which are obtained from a regression of excess returns on durable consumption growth for the 32-year period after the demise of Brenton-Woods in 1971. As can be seen, the consumption risk is generally increasing in the interest rate, with a difference in consumption betas of the first and seventh portfolios of about 160 basis points. The remaining rows in Table 6 report the estimated standard errors and critical values for the six different tests of the regression coefficients. Numbers in brackets are estimated standard errors. Numbers in italics represent alternative critical values which can be used in the hypothesis tests. Significance at the 5% level is denoted by an asterisk.

When the Newey–West estimator is used to estimate the standard errors, only three portfolios exhibit risk parameters that are significantly different from zero. However, hypothesis rejection varies depending on which standard error estimator is employed. For example, the hac, hac^{KV}, and $\text{hac}^{\text{KV}n}$ tests all reject the hypothesis that β is equal to zero for the third portfolio. The hac^{KV} and $\text{hac}^{\text{KV}n}$ tests reject the hypothesis even though the test critical values have increased from 1.96 to 2.0794 and 4.813, respectively. However, the pw, cho, and cho^{R} tests fail to reject the hypothesis as the estimated standard errors are significantly larger than the Newey–West standard errors.

For portfolios 4 and 7, the hac and hac^{KV} tests continue to reject the null, but the other tests show some switching in test rejection across portfolios. Moving to portfolio 4, the $\text{hac}^{\text{KV}n}$ tests no longer rejects the null hypothesis, but the pw and cho tests now do. Last, in the seventh portfolio, every testing procedure, excepting the cho^{R} test, rejects the null hypothesis at the 5% level. Interestingly, the cho^{R} test fails to reject the null hypothesis for any of the eight portfolios.

Although it is impossible to know exactly what causes the test rejections to vary across portfolios – or for that matter whether the tests are properly or improperly rejecting the null – the simulation results in the previous section suggest that heteroskedasticity may play a role. To further examine the effect of heteroskedasticity, we test the null hypothesis of homoskedasticity in portfolios 3, 4, and 7 using the White heteroskedasticity test. Specifically, we regress the squared OLS residuals from each model on a constant, the consumption factor, f_t, and the consumption factor squared, then test the hypothesis that the coefficients on the factors are jointly equal to zero.[25] For portfolio 3, the χ^2-test statistic with two degrees of freedom is 1.613 with a p-value of 0.446. For portfolios 4 and 7, the tests statistics are 3.242 and 8.310 and the p-values are 0.198 and 0.016, respectively. There appears to be little evidence of heteroskedasticity for portfolio 3, slight evidence of heteroskedasticity in portfolio 4, and substantial evidence of heteroskedasticity in

Table 6. Estimation of factor betas for eight portfolios sorted on interest rates (1971–2002).

	Portfolio							
Coef. estimate	1	2	3	4	5	6	7	8
β	0.5367	0.7857	1.2881	2.0321	1.2249	1.3590	2.1827	0.8447
				[Standard error estimate] *Test critical value*				
hac	[0.7177] *1.96*	[0.5437] *1.96*	[0.5503]* *1.96*	[0.7371]* *1.96*	[0.7453] *1.96*	[0.9185] *1.96*	[0.7997]* *1.96*	[0.8608] *1.96*
hac^{KV}	[0.7177] *2.1471*	[0.5437] *2.1471*	[0.5503]* *2.1471*	[0.7371]* *2.0532*	[0.7453] *2.1471*	[0.9185] *2.0532*	[0.7997]* *2.0532*	[0.8608] *2.2416*
pw	[0.8108] *1.96*	[0.5796] *1.96*	[0.6712] *1.96*	[0.7762]* *1.96*	[0.8125] *1.96*	[0.9431] *1.96*	[0.7532]* *1.96*	[0.8516] *1.96*
hac^{KVn}	[0.5759] *4.813*	[0.4333] *4.813*	[0.2016]* *4.813*	[0.4434] *4.813*	[0.4511] *4.813*	[0.4013] *4.813*	[0.3185]* *4.813*	[0.4286] *4.813*
cho	[0.8597] *1.96*	[0.7086] *1.96*	[0.8178] *1.96*	[0.9615]* *1.96*	[0.8409] *1.96*	[0.9694] *1.96*	[1.0425]* *1.96*	[1.5109] *1.96*
cho^{R}	[0.8597] *2.2717*	[0.7086] *2.2636*	[0.8178] *2.2417*	[0.9615] *2.2678*	[0.8409] *2.3921*	[0.9694] *2.2654*	[1.0425] *2.2422*	[1.5109] *2.2476*

portfolio 7. The results in the previous section suggest that the cho and cho$^{\mathrm{R}}$ tests are likely most accurate in portfolio 3, indicating the hac, hac$^{\mathrm{KV}}$, and hac$^{\mathrm{KV}n}$ tests may falsely reject the null. On the other hand, the cho and cho$^{\mathrm{R}}$ tests are most likely to overreject the null hypothesis in portfolio 7, when heteroskedasticity is strongest. Indeed, the cho test rejects the null hypothesis in portfolio 7. However, when the conditionally homoskedastic estimator is used in conjunction with the Rothenberg critical value, the test fails to reject the hypothesis that β is equal to zero.

5. Conclusion

The simulation results of this paper suggest that when sample sizes are small, modelling the heterogeneity of a process is secondary to accounting for dependence. We find that a conditionally homoskedastic covariance matrix estimator (when used in conjunction with Rothenberg's second-order critical value adjustment) improves test size with only a minimal loss in test power, even when the data manifest significant amounts of heteroskedasticity. In some specifications, the size inflation was cut by nearly 40% over the traditional HAC test. Much of this size gain can be attributed to the manner in which the second-order theory adjusts the critical value according to the degree of correlation found in the data. Strong dependence usually leads to estimated standard errors that are too small, resulting in tests that reject the null hypothesis too often. Rothenberg's second-order critical value refinements serve to dampen this effect by producing larger critical values when standard errors are small, and more moderate critical values when standard errors are large.

In addition to improved small-sample test size, the conditionally homoskedastic estimator offers improvements over traditional HAC estimators in both computational ease and testing continuity. While the second-order critical value calculation is slightly complicated, the process is completely data dependent and requires no user choices in implementation. Conversely, traditional HAC estimators require the user to specify a weighting kernel, prewhitening filter, and a bandwidth selection procedure, which previous authors have shown can have a significant impact on the performance of the estimator.

To be sure, the conditionally homoskedastic covariance estimator does not perform the best under all heteroskedastic conditions. As noted by Rothenberg, statistical inference is especially problematic when the error variance is highly correlated with the regressor of interest. Care must be taken when heteroskedasticity is caused primarily by the regressor whose coefficient is under test. However, the adjusted critical values and estimator that we propose deliver a testing procedure with substantial gains in controlling size, at little cost in power.

Acknowledgements

We thank the members of the Econometrics Research Group at the University of California, Santa Barbara for helpful comments.

Notes

1. Other variance estimators have also been proposed such as the vector autoregression heteroskedasticity and autocorrelation consistent (VARHAC) estimator [6] or the bootstrap estimator [7–10]. The simulation results of Den Haan and Levin [6,11] show that the VARHAC estimator performs comparably to the Andrews and Monahan prewhitened estimator, and Kiefer and Vogelsang [2] show that the block bootstrap performs comparably to tests employing the fixed-bandwidth asymptotic critical values.
2. Rothenberg assumes that the regressors are strictly exogenous, while Newey–West, Andrews–Monahan and Kiefer–Vogelsang assume that the regressors are only weakly exogenous. The simulations in Erb and Steigerwald [13] study the relative performance of the Rothenberg correction under both strict and weak exogeneity.

3. We set ε_0 and η_{k0} equal to zero in the simulations and discard the first 50 observations to remove any influence from initial values.
4. Data were also simulated according to a moving-average specification, under which the serial correlation is of limited duration and conditional heteroskedasticity plays a correspondingly larger role. The results were similar to those found in the autoregressive specifications and can be found in Erb and Steigerwald [13].
5. The values of ζ have been chosen to ensure that the unconditional variance of U_t is the same in all specifications. Andrews and Monahan also study this specification, albeit without reference to the findings in Rothenberg.
6. To see that this model brings conditional heteroskedasticity, consider the case in which $\zeta = (0, 1, 0, 0, 0)'$. Then

$$E(U_t U_{t-1} \mid X) = E(|X_{2t}|\tilde{U}_t \cdot |X_{2t-1}|\tilde{U}_{t-1} \mid X),$$

and the covariance conditional on X is $\rho_u/(1-\rho_u^2)|X_{2t}||X_{2t-1}|$.
7. Not considered here are the classic OLS variance estimator under the assumption of i.i.d. errors and the parametric, AR(1) variance estimator. Simulation results for these estimators can be found in Erb and Steigerwald [13].
8. We use $n-5$, rather than n, as the divisor because the degrees-of-freedom calculation is likely to be used when the sample size is small, as is the case in our simulations where $n = 50$.
9. To ensure the matrix $I - \sum_s A_s$ is not too close to singularity, we restrict the eigenvalues of $\sum_s \hat{A}_s$ to be no larger than 0.97 in absolute value. See Andrews and Monahan [5, p. 957] for the details.
10. This estimator, which includes prewhitening, retains many of the asymptotic properties of $\hat{J}_{\text{hac}}$ including the rate of convergence.
11. Our finite-sample results are designed to guide researchers with moderate sample sizes, in which size inflation is known to be a problem. Although $\hat{J}_{\text{cho}}$ is not a consistent estimator of J under conditional heteroskedasticity, a consistent estimator is easily obtained by switching from $\hat{J}_{\text{cho}}$ to $\hat{J}_{\text{hac}}$ for large sample sizes (say $n > 1000$).
12. A similar limit theory is developed by Phillips et al. [14].
13. Our goal is to evaluate the accuracy of the Kiefer–Vogelsang limit theory under both 'large' and 'small' specifications of the bandwidth parameter, and we use the Andrews method to select the 'small' bandwidth. Technically, the limit theory will not hold if the Andrews method is employed, as the Andrews procedure gives a bandwidth value that is $o(n)$ and Kiefer and Vogelsang assume that the bandwidth parameter is $O(n)$. However, we find in our simulations that the Kiefer–Vogelsang testing procedure performs substantially better (in terms of test size) when the Andrews procedure is implemented for each replication, as opposed to using some fixed value for the bandwidth, say $m/n = 0.2$, for all replications. For this reason, we report the simulation results using the Andrews method.
14. Unfortunately, it is not clear whether the asymptotic theory of Kiefer and Vogelsang extends to tests employing the prewhitened HAC estimator. The Kiefer–Vogelsang critical values depend on the ratio of the bandwidth to the sample size (m/n) in the HAC estimator. Although the prewhitened variance estimator also requires a specified bandwidth in the second stage, it is typically much smaller than the bandwidth without the first-stage prewhitening (as much of the dependence has already been removed). Although the empirical size may be improved in some cases by prewhitening, the limit theory may not hold in all cases. For example, Vogelsang and Franses [15] employ prewhitening with the alternative critical values and find (at least in one simulation specification) that, 'over-rejections can be a problem with moderately persistent data and prewhitening may not improve matters' (p. 15).
15. The precise form of f is detailed in the appendix.
16. Rothenberg considers covariance estimators of the form $1/(n-j)\sum \hat{U}_t\hat{U}_{t-j}$, where j corrects for the number of observations lost due to the lag length. To ensure a positive semi-definite estimator, we replace the factor $1/(n-j)$ with $1/(n-5)$ [16].
17. As the table makes clear, the conditionally homoskedastic variance estimator can only be recommended in conjunction with Rothenberg's second-order critical value adjustment when the data are heteroskedastic. Although the cho test has better size than the hac test if the serial correlation is high, it rarely exhibits smaller size than the pw or hac$^{\text{KV}}$ tests.
18. Intuition would suggest that for a fixed value of ρ_U the performance of the cho and cho$^{\text{R}}$ tests should deteriorate as ρ_X increases. However, ρ_X also adds to the overall pattern of serial correlation in the regressors as well as the errors. For this reason, the cho and cho$^{\text{R}}$ tests show size gains over other HAC tests when ρ_U is large, even when ρ_X is large.
19. For example, if the upper bound is set at 6, then the distance between each element of ψ is 6/10, and $\psi = \{0,\ 0.6,\ 1.2,\ 1.8, \ldots, 6\}$.
20. Note that the range of values along the β_2-axis differs for alternative values of the serial correlation parameters. All tests are showing substantial increases in power as the correlation in the data falls.
21. Researchers often prefer 'automatic' bandwidth procedures as they simplify the estimation process and provide a level of uniformity across different projects. However, researchers must still make decisions regarding the specifications which underpin all mechanical bandwidth methods, such as objective functions and data dependence. For example, the Andrews bandwidth procedure is a function of J, and therefore, a preliminary estimate of J is necessary to compute the bandwidth to be used in the estimation of $\hat{J}_{\text{hac}}$. Typically, the AR(1) model is used as a preliminary estimate of J, but alternative models will lead to alternative bandwidths.
22. There is no data-dependent method of choosing an 'optimal' bandwidth for the hac$^{\text{KV}}$ test. Phillips, Sun, and Jin propose a data dependent rule for their test that minimizes a weighted sum of type I and type II errors, for which Kiefer and Vogelsang conjecture can be extended to their test. However, in place of selecting the bandwidth, the researcher is now left to choose the proper weights for the type I and type II errors.

[23] To ensure the estimator accurately accounts for the serial correlation, we impose a minimum bandwidth of 5. Allowing for bandwidths less than 5 further inflates the rejection probabilities.

[24] Not surprisingly, when the errors truly are homoskedastic, the cho$^{\mathrm{R}}$ test outperforms all other robust tests in terms of finite-sample empirical size, as it exploits the homogeneity in the data. These results can be found in Erb and Steigerwald [13].

[25] A heteroskedasticity pretest would undoubtedly alter the test size and we do not recommend choosing among the testing procedures in this manner.

Appendix. Calculation of second-order critical values

We are interested in testing the single restriction hypothesis $H_0 : c'\beta = c'\beta_0$. Under the assumption of conditional homoskedasticity, we form the test statistic

$$t_{\text{cho}} = \frac{c'\sqrt{n}(B-\beta_0)}{[c'(n^{-1}X'X)^{-1}\hat{J}_{\text{cho}}(n^{-1}X'X)^{-1}c]^{1/2}}.$$

Using Edgeworth expansions, Rothenberg shows that the second-order-adjusted critical value of this test statistic, defined as $\Pr(t_{\text{cho}} > \text{cv}_\alpha^{\mathrm{R}}) = \alpha$, can be expressed as

$$\text{cv}_\alpha^{\mathrm{R}} = Z_\alpha\left(1 + \frac{1/4(1+Z_\alpha^2)\hat{V}_W - \hat{a}(Z_\alpha^2 - 1) - \hat{b}}{2n}\right)$$

where Z_α is the corresponding α critical value from the standard normal distribution. The formula for $\hat{V}_W$ comes from rewriting the test statistic t_{cho} as

$$t_{\text{cho}} = \frac{T_{\text{cho}}}{(1 + W/\sqrt{n})^{1/2}}.$$

Here, T_{cho} is the test statistic formed using the true value of J_{cho}, and by definition,

$$T_{\text{cho}} = \frac{c'(B-\beta_0)}{[c'(X'X)^{-1}J_{\text{cho}}(X'X)^{-1}c]^{1/2}}$$

and

$$W = \sqrt{n}\frac{c'(X'X)^{-1}(n\hat{J}_{\text{cho}} - J_{\text{cho}})(X'X)^{-1}c}{c'(X'X)^{-1}J_{\text{cho}}(X'X)^{-1}c}.$$

If the regression errors are known to be conditionally homoskedastic and stationary with unknown autocovariances, Rothenberg (p. 1006) derives the specific form of W as well as its variance. The estimator of the variance of W is

$$\hat{V}_W = \frac{2\sum_k\left(\sum_j r_j\hat{\delta}_{j+k}\right)^2}{\left(\sum_k r_k\hat{\delta}_k\right)^2}, \qquad \begin{array}{l} k = -(n-1),\ldots,0,\ldots(n-1) \\ j = -(n-1),\ldots,0,\ldots(n-1) \end{array}$$

where $\hat{\delta}_j$ is the jth sample autocovariance as defined in Section 2, and

$$r_j = n^{-1}\sum_{t=1}^{n-|j|} x_t x_{t+|j|} \quad j = -(n-1),\ldots,0,\ldots,n-1$$

with $x_t = nX(X'X)^{-1}c$.

The parameters $\hat{a}$ and $\hat{b}$ are defined as

$$\hat{a} = \frac{\sum_k r_k\bar{r}_k}{\sum_k r_k\hat{\delta}_k} \quad \text{and} \quad \hat{b} = \frac{\sum_k r_k\bar{q}_{kk}}{\sum_k r_k\hat{\delta}_k}, \quad k = -(n-1),\ldots,(n-1)$$

where $\bar{r}_k = (n-k)^{-1}\sum_t \hat{z}_t\hat{z}_{t+k}$, and

$$\begin{aligned}\bar{q}_{kk} = {} & \text{trace}[(X'X)^{-1}(X'X_{-k})(X'X)^{-1}(n\hat{J}_{\text{cho}})] \\ & - 2\times\text{trace}[(X'X)^{-1}X'\hat{\Delta}X_{-k}].\end{aligned}$$

In the equation for $\bar{r}_k$, $\hat{z}_t$ is defined as the tth element of the $n \times 1$ vector

$$\hat{z} = \frac{M\hat{\Delta}x}{\sqrt{n^{-1}x'\hat{\Delta}x}},$$

where M is the $n \times n$ matrix $I_n - X(X'X)^{-1}X'$, and $\hat{\Delta}$ the $n \times n$ estimated covariance matrix for U with (s, t) element equal to $\hat{\delta}_{|s-t|}$. The lagged cross product matrices $X'X_{-k}$ and $X'\hat{\Delta}X_{-k}$ are formed by summing over the $n - |k|$ common observations.

References

[1] W. Newey and K. West, *A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix,* Econometrica 55 (1987), pp. 703–708.

[2] N. Kiefer and T. Vogelsang, *A new asymptotic theory for heteroskedasticity-autocorrelation robust tests*, Econom. Theory 21 (2005), pp. 1130–1164.

[3] H. White and I. Domowitz, *Nonlinear regression with dependent observations*, Econometrica 52 (1984), pp. 143–162.

[4] D. Andrews, *Heteroskedasticity and autocorrelation consistent covariance matrix estimation*, Econometrica 59 (1991), pp. 817–858.

[5] D. Andrews and J. Monahan, *An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator*, Econometrica 60 (1992), pp. 953–966.

[6] W. Den Haan and A. Levin, *Robust covariance matrix estimation with data-dependent VAR prewhitening order*, Technical Working Paper No. 255, National Bureau of Economic Research, 2000.

[7] F. Götze and H. Künsch, *Second-order correctness of the blockwise bootstrap for stationary observations*, Ann. Statist. 24 (1996), pp. 1914–1933.

[8] P. Hall and J. Horowitz, *Bootstrap estimators for tests based on generalized method of moments estimators*, Econometrica 64 (1996), pp. 891–916.

[9] B. Huitema, J. McKean, and S. McKnight, *A double bootstrap method to analyze linear models with autoregressive error terms*, Psychol. Methods 5 (2000), pp. 87–101.

[10] S. Goncalves and T. Vogelsang, *Block bootstrap puzzles in HAC robust testing: The sophistication of the naive bootstrap*, Working Paper, Department of Economics, Cornell University, 2004.

[11] W. Den Haan and A. Levin, *A practitioner's guide to robust covariance matrix estimation*, Technical Working Paper No. 197, National Bureau of Economic Research, 1996.

[12] T. Rothenberg, *Approximate power functions for some robust tests of regression coefficients*, Econometrica 56 (1988), pp. 997–1019.

[13] J. Erb and D. Steigerwald, *Accurately sized test statistics with misspecified conditional homoskedasticity*, working paper, Department of Economics, University of California, Santa Barbara, 2009.

[14] P. Phillips, Y. Sun, and S. Jin, *Spectral density estimation and robust hypothesis testing using steep origin kernels without truncation*, Internat. Econom. Rev. 47 (2006), pp. 837–894.

[15] T. Vogelsang and P. Franses, *Testing for common deterministic trend slopes*, J. Econom. 126 (2005), pp. 1–24.

[16] A. McLeod and C. Jimenez, *Nonnegative definiteness of the sample autocovariance function*, Amer. Statist. 38 (1984), pp. 297–298.

[17] H. Lustig and A. Verdelhan, *The cross section of foreign currency risk premia and consumption growth risk*, Am. Econ. Rev. 97 (2007), pp. 89–117.