經濟與管理論叢(Journal of Economics and Management), 2009, Vol. 5, No. 2, 219-240

Assessing the Accuracy of Event Forecasts

Ching-Chuan Tsong*

Department of Economics, National Chi Nan University, Taiwan

Event forecasts, often generated from estimated econometric models, comprise a binary time series. In empirical finance, the market timing test proposed by Henricksson and Merton (1981) is probably the most popular method to assess the accuracy of these forecasts. Unfortunately, event forecasts and/or realizations are serially correlated, violating the independent identical distributed (IID) assumption. Consequently, the market timing test has an inflated size that can lead to doubtful empirical results. We find that the heteroskedasticity-autocorrelation (HAC) robust *t*-test with fixed-*b* asymptotics in Kiefer and Vogelsang (2005) and with the empirical distribution obtained using the naive block bootstrap can overcome this problem. As compared to several extant testing methods, simulation results reveal that the empirical size of these two testing procedures is quite close to the nominal size in finite samples. An empirical study is performed to demonstrate the usefulness of the naive block bootstrap.

Keywords: naive block bootstrap, HAC robust test, market timing test
JEL classification: C12, C53

1 Introduction

Used in decision-making, forecasting has always been an important topic in economics and finance. Government institutions make policy decisions via forecasts of certain economic variables, and firms depend on forecasting to guide their investment decisions. While real-valued point forecasts are popular in the academic

Received November 15, 2008, revised February 16, 2009, accepted June 2, 2009.
*Correspondence to: Department of Economics, National Chi Nan University, Nantou 545, Taiwan. Tel: (88649) 291-0960 ext. 4662; Fax: (88649) 291-4435; E-mail: tcc126@ncnu.edu.tw. Tsong gratefully acknowledges the anonymous referee's helpful comments and thanks Pau-Yu Lu for undertaking data collection.

area, financial practitioners rely more on direction-of-change forecasts for their investments. For instance, if stock market returns will be higher than the yield of treasury bills, investors should reallocate their funds from the bond market to the stock market to gain more profits. Engel (1994) predicts the direction of change of 18 exchange rates by using the Markov switching model. A challenge that practitioners face is whether or not the event forecasts are accurate. One answer to this issue is to develop an out-of-sample test for evaluating the event forecasts on hand. This financial literature was pioneered by Henricksson and Merton (1981) who proposed the market timing test (henceforth, the HM test). Since then, the HM test has been used in a wide variety of applications.

Although the HM test was proposed for this purpose, its use may not be appropriate in time-series situations. Event forecasts—obtained from truncating certain predetermined value of point forecasts generated by estimated econometric models—may be serially correlated. Similarly, event realizations generated by an unknown process are possibly autocorrelated as well. The serial correlation of event forecasts and/or realizations violates the maintained IID assumption for the HM test. In this paper, we first show through a simulation that the IID assumption is crucial for the validity of the HM test. Without this assumption, this test has severe size distortions that can falsely reject the null hypothesis and lead to questionable empirical results. Therefore, a robust test that can account for autocorrelated event forecasts and realizations is extremely important to portfolio managers. However, to our knowledge, the issue has received—and continues to receive—little attention in the financial literature thus far.

The conventional HAC robust *t*-test (e.g., Newey and West, 1987) in a regression framework accommodates serially correlated disturbances, and therefore, it should be a good starting point for getting a robust test. Under the assumption that as the sample size (n) grows, the number (M) of sample autocovariances becomes infinite and the fraction (M/n) of sample autocovariances for the variance estimator tends to zero, the HAC estimator is consistent for asymptotic variance. Hence, the asymptotics of the HAC robust *t*-test can be derived as though the variance were known. While the asymptotics follows the standard normal distribution, the HAC robust *t*-test still has a tendency to over reject the null hypothesis in finite samples (e.g., Andrews, 1991). Therefore, the HAC robust *t*-test with the standard normal

distribution is only a partial solution to the over-sized issue.

In practice, however, given a particular data set, a practitioner uses some positive fraction of autocovariances to estimate the asymptotic variance. This implies that M/n should be a positive number less than or equal to unity. Based on this fact, Kiefer and Vogelsang (2005) derived a brand-new asymptotic theory for the asymptotic variance estimator under the assumption of $M = bn$, where $b \in (0,1]$. While the HAC variance estimator is no longer consistent in this case, its asymptotic distribution is proportionate to the unknown asymptotic variance and depends on the kernel and *b*.

In addition, the HAC robust *t*-test has pivotal asymptotics (henceforth, fixed-*b* asymptotics) that, however, depends on the kernel and *b*. This differs from the standard normal distribution wherein the effects of kernel and bandwidth are not involved. Derived under the assumption of $M = bn$, where $b \in (0,1]$, which reflects the situations in empirical applications, the fixed-*b* asymptotics is a more accurate approximation to the sampling distribution of the HAC robust *t*-test than standard normal distribution. Therefore, with the critical value from the fixed-*b* asymptotics, over-rejections can be reduced remarkably.

Bootstrap is an alternative approximation to the sampling distribution of a test statistic. With an appropriate resampling procedure, bootstrap is an effective method to reduce the size distortions of a test statistic (e.g., Davison and Hall, 1993; Lahiri, 1996; Andrew, 2002). Based on this concept, Gonçalves and Vogelsang (2006) proposed the naive block bootstrap, where the formulas used on the bootstrap sample and the original data to compute the test are identical. They showed that the naive block bootstrap has the same large-sample distribution as the fixed-*b* asymptotics. Most importantly, evidence from our simulation shows that as compared with the fixed-*b* asymptotics, the empirical distribution obtained from the naive block bootstrap is a more accurate approximation to the sampling distribution of the HAC robust *t*-test in finite samples. This implies that the naive block bootstrap can deliver a more accurate size than fixed-*b* asymptotics in small samples even when event forecasts and/or realizations are serially correlated. In this paper, we rely on the naive block bootstrap to deal with the over-rejections of the HAC robust *t*-test used to assess the accuracy of event forecasts.

The remainder of the paper is organized as follows. Section 2 reviews some

extant tests and the HAC robust *t*-test for evaluating the accuracy of event forecasts. Section 3 reports the simulation results showing that the naive block bootstrap is a promising approach to overcome the over-sized problem. In Section 4, we provide an illustrative application. Section 5 summarizes the paper and offers some concluding remarks.

2 Evaluating Event Forecasts

Suppose that $\{Y_t\}_{t=1}^n$ is a binary stochastic process denoting out-of-sample event forecasts from a certain econometric model, and $\{X_t\}_{t=1}^n$ is the corresponding stochastic process of event realizations. Under the maintained assumption that both $\{Y_t\}_{t=1}^n$ and $\{X_t\}_{t=1}^n$ are individually IID processes, Henriksson and Merton (1981) test the null hypothesis of no timing ability, that is,

$$H_0 : P(Y_t = 0 \mid X_t = 0) + P(Y_t = 1 \mid X_t = 1) = 1 . \tag{1}$$

This is a test of contemporary independence between $\{Y_t\}_{t=1}^n$ and $\{X_t\}_{t=1}^n$. Under the null hypothesis, they showed that the test statistic of $\#\{Y_t = 0 \mid X_t = 0\}$ has a hypergeometric distribution and can be written as

$$P(\#\{Y_t = 0 \mid X_t = 0\} = k) = \frac{C_k^{N_1} C_{m-k}^{N_2}}{C_m^n}, \tag{2}$$

where N_1 and N_2 denote the number of $X_t = 0$ and $X_t = 1$, respectively, and m denotes the number of $Y_t = 0$. Testing the null hypothesis of Eq. (1) is straightforward with the critical value obtained from Eq. (2) in a small sample. For large samples, however, the computation of factorials can be quite tedious. Fortunately, for large samples, the hypergeometric distribution can be accurately approximated by the normal distribution with mean μ and variance σ^2 described as

$$\mu = \frac{mN_1}{n} \quad \text{and} \tag{3}$$

$$\sigma^2 = \frac{n_1 N_1 (n - N_1)(n - m)}{n^2 (n-1)}, \tag{4}$$

where n_1 is the number of $Y_t = 0$ given $X_t = 0$, and other parameters are defined

as above.

In the framework of Henriksson and Merton (1981), independence between Y_t and X_t is tested, regarding IID stochastic processes of $\{Y_t\}_{t=1}^n$ and $\{X_t\}_{t=1}^n$ as the maintained assumption. Under this setting, the conventional chi-square test of independence as well as the *t*-test in the regression framework can be used to test the null hypothesis. Since Y_t can be viewed as a binary dependent variable, the logit and probit models are also appropriate for this problem.

Unfortunately, event forecasts, obtained from truncating a certain predetermined value of the point forecasts generated by estimated econometric models, may be serially correlated. Similarly, event realizations have a serial correlation as well. Therefore, the IID assumption is violated for all the abovementioned test procedures. As a result, these tests suffer from severe size distortions, falsely rejecting the null hypothesis too often.[1]

To deal with the over-rejections, the testing procedure should accommodate the autocorrelation of event forecasts and realizations. Three testing procedures, including the HAC robust *t*-test with the standard normal distribution, the HAC robust *t*-test with fixed-*b* asymptotics, and the naive block bootstrap are described briefly below. Breen *et al*. (1989) pointed out that event forecasts evaluation can be proceeded with a *t*-test for $\beta = 0$ in the linear regression:

$$Y_t = \alpha + \beta X_t + u_t, \quad t = 1, 2, \ldots, n. \tag{5}$$

For ease of exposition, rewrite Eq. (5) as

$$y_t = x_t'\gamma + u_t, \quad t = 1, 2, \ldots, n, \tag{6}$$

where $y_t = Y_t$, $x_t = (1, X_t)'$ and $\gamma = (\alpha, \beta)'$. Under some regularity conditions, it is straightforward that

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N\left(0, Q^{-1}\Omega Q^{-1}\right), \tag{7}$$

where $\hat{\gamma}$ is the least squares (LS) estimator for γ, $Q = p\lim n^{-1}\sum_{t=1}^n x_t x_t'$ and Ω denotes the long-run variance of $v_t = x_t u_t$. Testing hypotheses about γ involves getting consistent estimators of $\hat{Q}^{-1}\hat{\Omega}\hat{Q}^{-1}$ for $Q^{-1}\Omega Q^{-1}$, following which the asymptotics free from the nuisance parameters can be constructed by using Eq. (7).

[1]Simulation results in the next section will confirm this finding.

Clearly, Q can be consistently estimated by $\hat{Q} = n^{-1}\sum_{t=1}^{n} x_t x_t'$. A consistent estimation of Ω, however, is more complicated, since v_t may be serially correlated with an unknown pattern. In the literature, a kernel-based consistent estimator for Ω is the most popular. It can be defined as

$$\hat{\Omega} = \sum_{j=-(n-1)}^{n-1} k(j/M)\hat{\Gamma}_j \,, \tag{8}$$

with

$$\hat{\Gamma}_j = \frac{1}{n}\sum_{t=j+1}^{n} \hat{v}_t \hat{v}_{t-j}' \quad \text{for} \quad j \geq 0 \,, \quad \hat{\Gamma}_j = \hat{\Gamma}_{-j}' \quad \text{for} \quad j < 0 \,, \tag{9}$$

where $k(x): \Re \to [-1,1]$ is an even kernel function satisfying $k(0) = 1$, $k(x)$ continuous at $x = 0$ and $\int_{-\infty}^{\infty} k(x)dx < \infty$; $\hat{v}_t = x_t \hat{u}_t$ and $\hat{u}_t = y_t - x_t'\hat{\gamma}$. Often, M denotes the bandwidth as $k(x) = 0$ for $|x| > 1$. With the conditions of $M \to \infty$ and $M/n \to 0$ as $n \to \infty$, $\hat{\Omega}$ is consistent for Ω. Based on Eq. (7) and the argument discussed above, the HAC robust *t*-test for the evaluation of event forecasts and its asymptotic distribution can be written as

$$t_{HAC} = \frac{\sqrt{n}R\hat{\gamma}}{\sqrt{R\hat{Q}^{-1}\hat{\Omega}\hat{Q}^{-1}R'}} \xrightarrow{d} N(0,1) \,, \tag{10}$$

where $R = (0,1)$. This result implies that with a suitable choice of bandwidth, the standard normal distribution can be served as an approximation to the sampling distribution of the HAC robust *t*-test, regardless of which kernel is used.

While the conditions of $M \to \infty$ and $M/n \to 0$ as $n \to \infty$ are essential to derive the asymptotics of t_{HAC}, they cannot be satisfied in practice. In reality, a practitioner is given a particular data set, and the fraction of sample autocovariances used to compute $\hat{\Omega}$ is always a positive number smaller than unity. Based on this fact, Kiefer and Vogelsang (2005) derived a brand-new asymptotic theory known as fixed-*b* asymptotics, for the HAC robust *t*-test under the condition that the bandwidth is set as a fixed ratio of the sample size. Important differences deserve to be stressed. In this scenario, $\hat{\Omega}$ is no longer a consistent estimator for Ω, but converges to a random matrix that is proportional to Ω. The HAC robust *t*-test computed in the usual manner has an asymptotic distribution that depends on the

kernel and the bandwidth through *b*, but is free from any nuisance parameter. Therefore, the critical values can be tabulated and served as the purpose for hypothesis testing. Most importantly, the fixed-*b* asymptotics can deliver a more accurate approximation than the standard normal distribution, since the value of *b* is greater than zero and less than or equal to unity in practice. As a result, the HAC robust *t*-test computed in the usual manner can effectively reduce size distortions with fixed-*b* asymptotics, as compared to that with its standard normal counterpart. Kiefer and Vogelsang (2002) argued that with the choice of $M = n$, i.e., $b = 1$ and the Bartlett kernel, the HAC robust *t*-test would have a good size and reasonable power performance. Hence, we follow their suggestion in the simulation experiments and empirical study described below.

The bootstrap is an alternative to asymptotic approximations. With an appropriate resample scheme, the bootstrap asymptotics can be a more accurate approximation to the sampling distribution of a test (e.g., Lahiri, 1996; Götze and Künsch, 1996; Park, 2003, among others). According to this idea, Gonçalves and Vogelsang (2006) proposed the naive bootstrap where the formula used to compute the test for the bootstrap sample is the same as that used for the original data. The naive bootstrap is briefly stated below. Let $w_t = (y_t, x_t')'$ be the vector that collects dependent and independent variables defined in Eq. (6) for each observation. Further, $B_i = (w_i, w_{i+1}, \ldots, w_{i+b-1})$ denotes the block of *b* consecutive observations starting from w_i, for $i = 1, 2, \ldots, n-b+1$. All the B_i together form $n-b+1$ overlapping blocks from the original sample w_t. With these $n-b+1$ overlapping blocks, the bootstrap sample $w_t^* = (y_t^*, x_t^*)'$ can be generated by randomly resampling n/b blocks with replacement and laying them end-to-end in the order in which they are sampled. Given this bootstrap sample, let $\hat{\gamma}^*$, $\hat{Q}^*$, and $\hat{\Omega}^*$ denote the bootstrap counterparts for $\hat{\gamma}$, $\hat{Q}$, and $\hat{\Omega}$, respectively, replacing w_t with w_t^*. Then, the naive block bootstrap HAC robust *t*-test is defined as

$$t^* = \frac{\sqrt{n}(R\hat{\gamma}^* - r^*)}{\sqrt{R\hat{Q}^{*-1}\hat{\Omega}^*\hat{Q}^{*-1}R'}}, \tag{11}$$

where $r^* = R\hat{\gamma}$ and $R = (0,1)$. Repeat the sampling NB times, and the computed NB values of t^* can be regarded as an empirical distribution function of the HAC robust *t*-test. Subsequently, make an inference based on the bootstrap critical value

from the empirical distribution function. Gonçalves and Vogelsang (2006) also showed that the naive block bootstrap has the same limiting distribution as the fixed-*b* asymptotics. Moreover, their simulation results suggest that with the appropriate choice of block length, the naive block bootstrap can deliver a more accurate approximation than the fixed-*b* asymptotics. This is promising for the naive block bootstrap to deal with the over-rejections of extant tests used to evaluate the accuracy of event forecasts.

3 Monte Carlo Evidence

3.1 Experimental Design

In this section, we investigate the finite-sample performance of various test statistics discussed in Section 2. These tests are categorized into two groups by their capability of capturing serial correlations. Conventional test statistics, including the HM test, the chi-square test of independence, the *t*-tests in linear regression and logit models, fall into the first group.[2] The second group includes HAC robust *t*-test with three types of distributions as an approximation to its sampling distribution, such as the standard normal distribution, fixed-*b* asymptotics and the empirical distribution obtained from the naive block bootstrap. To generate binary series X_t and Y_t, let the data-generating process (DGP) be:

$$X_t = \begin{cases} 1 & \text{if } u_t^x > 0 \\ 0 & \text{if } u_t^x \le 0 \end{cases}, \tag{12}$$

and

$$Y_t = \begin{cases} 1 & \text{if } u_t^y > 0 \\ 0 & \text{if } u_t^y \le 0 \end{cases}, \tag{13}$$

where

$$u_t^x = \rho_x u_{t-1}^x + \varepsilon_t^x \tag{14}$$

$$u_t^y = \rho_y u_{t-1}^y + \varepsilon_t^y \tag{15}$$

[2]Since the *t*-test in the probit model yields similar results as in the logit model, we omit it for brevity.

with

$$\begin{bmatrix} \varepsilon_t^x \\ \varepsilon_t^y \end{bmatrix} \overset{IID}{\sim} N\left[0, \begin{pmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{pmatrix}\right] \tag{16}$$

and initials $u_0^x = u_0^y = 0$. Obviously, if $\rho_x \neq 0$, then u_t^x in Eq. (14) is an autocorrelated series, which leads X_t to a serially correlated series. Similarly, if $\rho_y \neq 0$, Y_t is autocorrelated. On the other hand, ρ_{xy} in Eq. (16) measures the contemporary correlation between ε_t^x and ε_t^y. With the definitions in Eq. (12) and Eq. (13), the value of ρ_{xy} also governs the contemporary correlation between X_t and Y_t. We consider the sample sizes of $n = 50$, 100, 200, and 1000; the contemporary correlation parameters of $\rho_{xy} = 0$, 0.2, 0.5, and 0.8; and the autoregressive parameters of ρ_x and ρ_y each equaling 0, 0.5, 0.8, and 0.9.

For all the tests except for the naive block bootstrap, the empirical sizes are computed when $\rho_{xy} = 0$; otherwise, the size-adjusted powers are calculated. For the naive block bootstrap, however, we compute the bootstrap sizes and powers corresponding respectively to $\rho_{xy} = 0$ and otherwise. Before implementing the naive block bootstrap, the block length must first be determined. Although Politis and White (2004) proposed an automatic method to select the block length for the block bootstrap, our simulation results show that it is futile in dealing with the over-sized problem.[3] Instead, we choose $b = \text{int}(n^{1/5})$, where int(.) denotes the integer part, since Hall and Jing (1996) pointed out that the optimal block length should be proportional to $n^{1/5}$ in this context. All the results are under a nominal size of 5%. We perform 5,000 Monte Carlo replications for the asymptotic tests. For the naive block bootstrap, on the other hand, the number of replications is 1,000 (= NB). Note that computing the HAC robust *t*-test involves choosing the kernel and bandwidth. As mentioned in Section 2, we choose the Bartlett kernel for the power concern. For the standard normal approximation, bandwidth M is set as $\text{int}(12(n/100)^{1/4})$, while for fixed-$b$ asymptotics, we choose $M = n$, as Kiefer and Vogelsang (2002) suggested.

[3]For the sake of concision, these results are omitted. They can be obtained from the author upon request.

3.2 Finite Sample Properties

The simulation results are collected in Tables 1 and 2. The results for the empirical size of various tests are reported in Table 1. As expected, all the tests except for t_{HAC} in the case of $n = 50$ have an empirical size close to the nominal size when $\rho_x = \rho_y = 0$. When ρ_x and ρ_y increase, all the tests in the first group denoted by HM, x^2, logit-*t* and regression-*t* have severe size distortions even in large samples. For example, the size of HM on $n = 1000$ inflates to 0.402 from 0.047 as both ρ_x and ρ_y increase to 0.9 from 0. On the other hand, the tests in the second group, including t_{HAC}, $t_{\text{fixed}-b}$, and t^*, have a reasonable size performance, except for t_{HAC} when ρ_x and ρ_y are large. This confirms the finding in Andrew (1991) that t_{HAC} has size distortions in finite samples when data are strongly serially correlated. Besides, the performance of t^* is better than that of the $t_{\text{fixed}-b}$ and t_{HAC} tests. Moreover, the results of $t_{\text{fixed}-b}$ are considerably better than those of t_{HAC}. For instance, they are 0.080, 0.113, and 0.204 for t^*, $t_{\text{fixed}-b}$, and t_{HAC}, respectively, when $\rho_x = \rho_y = 0.9$ and $n = 100$. These facts indicate that the consideration of serial correlation has a positive influence on size performance, as in the tests in the second group. Further, fixed-*b* asymptotics is a more accurate approximation to the sampling distribution of the HAC robust *t*-test than the standard normal distribution. Moreover, the naive block bootstrap with the chosen block length can offer a more accurate approximation than fixed-*b* asymptotics.

The results for empirical power are shown in Table 2. The figures for t^* are bootstrap power; otherwise, they are size-adjusted power. Undoubtedly, all power increases occur with a larger sample size *n*. For any given values of ρ_x and ρ_y, as expected, all the tests have a higher power with a larger value of ρ_{xy}. The power performance would be damaged by the autocorrelation of X_t and/or Y_t. For example, for the HM test, the power declines to 0.562 from 0.952 when ρ_y is up to 0.9 from 0.5, given $\rho_{xy} = 0.2$, $\rho_x = 0.5$, and $n = 1000$. Clearly, all the tests except for $t_{\text{fixed}-b}$ and t^* have comparable power. As compared with t_{HAC}, the power is lower for $t_{\text{fixed}-b}$ due to the longer bandwidth. However, the differences can be neglected when the value of ρ_{xy} becomes large. Interestingly, for some cases (e.g., $\rho_x = \rho_y = 0.9$), the t^* test shows a power gain over its counterpart of $t_{\text{fixed}-b}$ to

some extent. An important point must be emphasized here. Generally speaking, the power of t^* is lower than the other tests, but it is feasible. In other words, the power reported for the tests except for t^* is infeasible because the size-adjusted critical values in finite samples are generally unknown in applications. However, the bootstrap power, in practice, is feasible for any given sample size. Therefore, due to the good size and feasible power properties, t^* is a suitable choice for empirical applications.

Table 1: Empirical Size Performance of the Tests

n	HM test	x^2 test	Logit-t	regresstion-t	t_{HAC}	$t_{\text{fixed}-b}$	t^*
			$\rho_{xy}=0 \quad \rho_x=0 \quad \rho_y=0$				
50	0.050	0.056	0.047	0.062	0.123	0.056	0.042
100	0.056	0.057	0.057	0.059	0.093	0.059	0.051
200	0.048	0.050	0.052	0.051	0.066	0.047	0.047
1000	0.047	0.047	0.066	0.048	0.054	0.050	0.050
			$\rho_{xy}=0 \quad \rho_x=0.5 \quad \rho_y=0.5$				
50	0.083	0.088	0.081	0.097	0.143	0.065	0.054
100	0.089	0.091	0.092	0.097	0.103	0.064	0.045
200	0.082	0.083	0.088	0.086	0.080	0.056	0.055
1000	0.075	0.076	0.105	0.076	0.055	0.045	0.055
			$\rho_{xy}=0 \quad \rho_x=0.5 \quad \rho_y=0.8$				
50	0.121	0.127	0.110	0.143	0.149	0.065	0.062
100	0.120	0.123	0.120	0.128	0.115	0.062	0.048
200	0.122	0.124	0.123	0.126	0.084	0.056	0.045
1000	0.124	0.125	0.145	0.125	0.059	0.053	0.062
			$\rho_{xy}=0 \quad \rho_x=0.5 \quad \rho_y=0.9$				
50	0.126	0.130	0.108	0.148	0.146	0.057	0.064
100	0.145	0.145	0.143	0.151	0.115	0.059	0.058
200	0.130	0.133	0.132	0.135	0.082	0.056	0.044
1000	0.140	0.140	0.152	0.141	0.059	0.052	0.062
			$\rho_{xy}=0 \quad \rho_x=0.8 \quad \rho_y=0.5$				
50	0.110	0.119	0.099	0.134	0.165	0.082	0.060
100	0.118	0.121	0.120	0.126	0.111	0.066	0.061
200	0.116	0.117	0.118	0.122	0.082	0.057	0.071
1000	0.116	0.117	0.149	0.117	0.046	0.046	0.045

Table 1: Empirical Size Performance of the Tests (continued)

n	HM test	x^2 test	Logit-t	regresstion-t	t_{HAC}	$t_{\text{fixed}-b}$	t^*
			$\rho_{xy}=0 \quad \rho_x=0.8 \quad \rho_y=0.8$				
50	0.198	0.204	0.174	0.224	0.206	0.099	0.072
100	0.229	0.232	0.228	0.240	0.148	0.081	0.052
200	0.227	0.227	0.230	0.231	0.105	0.065	0.065
1000	0.237	0.238	0.251	0.239	0.072	0.053	0.055
			$\rho_{xy}=0 \quad \rho_x=0.8 \quad \rho_y=0.9$				
50	0.233	0.235	0.194	0.255	0.205	0.093	0.080
100	0.282	0.285	0.276	0.293	0.159	0.082	0.055
200	0.286	0.287	0.287	0.292	0.113	0.067	0.060
1000	0.293	0.293	0.304	0.294	0.069	0.054	0.049
			$\rho_{xy}=0 \quad \rho_x=0.9 \quad \rho_y=0.5$				
50	0.121	0.126	0.104	0.144	0.213	0.125	0.101
100	0.135	0.137	0.133	0.144	0.128	0.074	0.073
200	0.145	0.146	0.150	0.152	0.091	0.064	0.066
1000	0.142	0.142	0.167	0.144	0.058	0.050	0.056
			$\rho_{xy}=0 \quad \rho_x=0.9 \quad \rho_y=0.8$				
50	0.220	0.225	0.181	0.249	0.260	0.137	0.104
100	0.273	0.276	0.266	0.285	0.171	0.097	0.070
200	0.281	0.283	0.285	0.286	0.117	0.073	0.075
1000	0.292	0.293	0.309	0.294	0.071	0.054	0.061
			$\rho_{xy}=0 \quad \rho_x=0.9 \quad \rho_y=0.9$				
50	0.271	0.275	0.208	0.299	0.275	0.137	0.123
100	0.352	0.355	0.336	0.363	0.204	0.113	0.080
200	0.363	0.364	0.364	0.371	0.146	0.079	0.078
1000	0.402	0.403	0.409	0.403	0.086	0.054	0.064

Notes: The DGP is described from Eq. (12) to Eq. (16). The HM test is the market timing test in Henricksson and Merton (1981). The x^2 test denotes the independence test. Logit-t and regression-t represent the conventional t-test in logit model and regression model, respectively. t_{HAC}, $t_{\text{fixed}-b}$, and t^* denote the HAC robust t-test with the standard normal distribution, fixed-b asymptotics, and empirical distribution from the naive block bootstrap, respectively. Refer to Section 2 for further details. The figures reported are the rejection frequencies at the 5% nominal significance level, based on 5,000 replications for the asymptotic tests, and 1,000 for t^* with 1,000 re-samples. The asymptotic critical value for the 5% level is 1.96 for all two-sided tests except for $t_{\text{fixed}-b}$. For $t_{\text{fixed}-b}$, the asymptotic critical value is 4.771 obtained from Table 1 in Kiefer and Vogelsang (2002).

Table 2: Empirical Power Performance of the Tests

n	HM test	x^2 test	Logit-t	regresstion-t	t_{HAC}	$t_{\text{fixed}-b}$	t^*
			$\rho_{xy}=0.2 \quad \rho_x=0 \quad \rho_y=0$				
50	0.148	0.143	0.144	0.149	0.128	0.119	0.117
100	0.249	0.237	0.250	0.249	0.221	0.180	0.192
200	0.461	0.454	0.461	0.461	0.434	0.353	0.314
1000	0.987	0.985	0.974	0.987	0.984	0.857	0.869
			$\rho_{xy}=0.2 \quad \rho_x=0.5 \quad \rho_y=0.5$				
50	0.125	0.125	0.124	0.112	0.108	0.085	0.114
100	0.197	0.191	0.189	0.198	0.171	0.148	0.165
200	0.365	0.363	0.355	0.365	0.334	0.260	0.259
1000	0.952	0.952	0.923	0.952	0.812	0.918	0.796
			$\rho_{xy}=0.2 \quad \rho_x=0.5 \quad \rho_y=0.8$				
50	0.087	0.085	0.091	0.086	0.086	0.080	0.097
100	0.132	0.134	0.133	0.132	0.133	0.112	0.124
200	0.235	0.234	0.239	0.235	0.212	0.186	0.173
1000	0.801	0.804	0.740	0.801	0.780	0.581	0.601
			$\rho_{xy}=0.2 \quad \rho_x=0.5 \quad \rho_y=0.9$				
50	0.064	0.071	0.071	0.064	0.067	0.066	0.084
100	0.098	0.098	0.095	0.098	0.089	0.081	0.101
200	0.130	0.140	0.131	0.130	0.143	0.124	0.112
1000	0.562	0.577	0.505	0.563	0.552	0.399	0.395
			$\rho_{xy}=0.2 \quad \rho_x=0.8 \quad \rho_y=0.5$				
50	0.095	0.095	0.093	0.095	0.089	0.087	0.091
100	0.139	0.139	0.135	0.139	0.131	0.113	0.123
200	0.217	0.219	0.218	0.217	0.218	0.172	0.185
1000	0.817	0.817	0.767	0.802	0.798	0.608	0.579

Table 2: Empirical Power Performance of the Tests (continued)

n	HM test	x^2 test	Logit-t	regresstion-t	t_{HAC}	$t_{\text{fixed}-b}$	t^*
			$\rho_{xy}=0.2 \quad \rho_x=0.8 \quad \rho_y=0.8$				
50	0.087	0.088	0.081	0.088	0.080	0.072	0.095
100	0.120	0.120	0.115	0.120	0.107	0.101	0.121
200	0.187	0.184	0.184	0.187	0.185	0.152	0.183
1000	0.702	0.699	0.675	0.702	0.683	0.509	0.507
			$\rho_{xy}=0.2 \quad \rho_x=0.8 \quad \rho_y=0.9$				
50	0.071	0.074	0.070	0.072	0.072	0.064	0.092
100	0.086	0.086	0.087	0.086	0.093	0.087	0.110
200	0.136	0.136	0.136	0.136	0.141	0.118	0.128
1000	0.515	0.539	0.483	0.515	0.528	0.377	0.364
			$\rho_{xy}=0.2 \quad \rho_x=0.9 \quad \rho_y=0.5$				
50	0.069	0.073	0.071	0.072	0.058	0.059	0.113
100	0.106	0.107	0.104	0.107	0.098	0.090	0.104
200	0.148	0.139	0.145	0.148	0.156	0.118	0.154
1000	0.571	0.563	0.539	0.571	0.570	0.461	0.404
			$\rho_{xy}=0.2 \quad \rho_x=0.9 \quad \rho_y=0.8$				
50	0.074	0.070	0.071	0.077	0.064	0.066	0.128
100	0.083	0.088	0.085	0.084	0.087	0.085	0.099
200	0.140	0.135	0.136	0.140	0.150	0.128	0.168
1000	0.531	0.533	0.512	0.531	0.517	0.436	0.383
			$\rho_{xy}=0.2 \quad \rho_x=0.9 \quad \rho_y=0.9$				
50	0.069	0.074	0.072	0.072	0.069	0.070	0.128
100	0.082	0.086	0.084	0.082	0.083	0.077	0.114
200	0.109	0.111	0.109	0.109	0.117	0.107	0.142
1000	0.426	0.436	0.400	0.426	0.426	0.299	0.316

Table 2: Empirical Power Performance of the Tests (continued)

n	HM test	x^2 test	Logit-t	regresstion-t	t_{HAC}	$t_{\text{fixed}-b}$	t^*
			$\rho_{xy}=0.5 \quad \rho_x=0 \quad \rho_y=0$				
50	0.673	0.644	0.666	0.6755	0.566	0.485	0.451
100	0.925	0.919	0.925	0.925	0.884	0.745	0.726
200	0.998	0.998	0.998	0.998	0.994	0.946	0.945
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			$\rho_{xy}=0.5 \quad \rho_x=0.5 \quad \rho_y=0.5$				
50	0.550	0.538	0.546	0.551	0.479	0.438	0.397
100	0.840	0.835	0.830	0.840	0.789	0.651	0.658
200	0.988	0.988	0.987	0.988	0.982	0.884	0.890
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			$\rho_{xy}=0.5 \quad \rho_x=0.5 \quad \rho_y=0.8$				
50	0.331	0.327	0.339	0.312	0.279	0.250	0.280
100	0.626	0.629	0.624	0.626	0.595	0.477	0.489
200	0.903	0.903	0.902	0.903	0.884	0.727	0.718
1000	1.000	1.000	1.000	1.000	0.991	1.000	0.999
			$\rho_{xy}=0.5 \quad \rho_x=0.5 \quad \rho_y=0.9$				
50	0.196	0.205	0.201	0.186	0.169	0.143	0.175
100	0.397	0.397	0.390	0.397	0.354	0.292	0.341
200	0.653	0.673	0.651	0.653	0.662	0.514	0.508
1000	1.000	1.000	0.996	1.000	1.000	0.999	0.981
			$\rho_{xy}=0.5 \quad \rho_x=0.8 \quad \rho_y=0.5$				
50	0.374	0.368	0.366	0.374	0.327	0.306	0.251
100	0.622	0.625	0.610	0.622	0.575	0.462	0.494
200	0.901	0.903	0.899	0.901	0.888	0.724	0.718
1000	1.000	1.000	1.000	1.000	1.000	0.999	0.999

Table 2: Empirical Power Performance of the Tests (continued)

n	HM test	x^2 test	Logit-t	regresstion-t	t_{HAC}	$t_{\text{fixed}-b}$	t^*
			$\rho_{xy}=0.5 \quad \rho_x=0.8 \quad \rho_y=0.8$				
50	0.314	0.314	0.289	0.316	0.278	0.234	0.265
100	0.525	0.526	0.518	0.525	0.476	0.406	0.431
200	0.844	0.838	0.837	0.844	0.827	0.676	0.672
1000	1.000	1.000	1.000	1.000	1.000	0.993	0.997
			$\rho_{xy}=0.5 \quad \rho_x=0.8 \quad \rho_y=0.9$				
50	0.232	0.229	0.205	0.224	0.198	0.152	0.186
100	0.370	0.375	0.359	0.370	0.347	0.297	0.367
200	0.665	0.666	0.660	0.665	0.646	0.497	0.542
1000	1.000	1.000	0.999	1.000	1.000	1.000	0.980
			$\rho_{xy}=0.5 \quad \rho_x=0.9 \quad \rho_y=0.5$				
50	0.225	0.226	0.218	0.227	0.137	0.157	0.196
100	0.395	0.397	0.389	0.395	0.364	0.305	0.327
200	0.677	0.662	0.671	0.677	0.670	0.503	0.531
1000	1.000	1.000	1.000	1.000	1.000	0.977	0.977
			$\rho_{xy}=0.5 \quad \rho_x=0.9 \quad \rho_y=0.8$				
50	0.243	0.227	0.214	0.244	0.188	0.151	0.259
100	0.376	0.388	0.378	0.376	0.346	0.316	0.373
200	0.686	0.681	0.678	0.686	0.674	0.515	0.564
1000	1.000	1.000	0.999	1.000	0.999	0.999	0.978
			$\rho_{xy}=0.5 \quad \rho_x=0.9 \quad \rho_y=0.9$				
50	0.201	0.201	0.175	0.196	0.187	0.176	0.211
100	0.310	0.319	0.302	0.310	0.290	0.250	0.342
200	0.550	0.554	0.546	0.550	0.536	0.447	0.481
1000	0.998	0.998	0.997	0.998	0.997	0.937	0.940

Table 2: Empirical Power Performance of the Tests (continued)

n	HM test	x^2 test	Logit-t	regresstion-t	t_{HAC}	$t_{\text{fixed}-b}$	t^*
			$\rho_{xy}=0.8 \quad \rho_x=0 \quad \rho_y=0$				
50	0.993	0.992	0.984	0.993	0.977	0.930	0.853
100	1.000	1.000	1.000	1.000	1.000	0.993	0.989
200	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			$\rho_{xy}=0.8 \quad \rho_x=0.5 \quad \rho_y=0.5$				
50	0.977	0.975	0.956	0.977	0.949	0.889	0.797
100	1.000	1.000	1.000	1.000	0.999	0.984	0.968
200	1.000	1.000	1.000	1.000	1.000	0.999	0.999
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			$\rho_{xy}=0.8 \quad \rho_x=0.5 \quad \rho_y=0.8$				
50	0.830	0.826	0.805	0.827	0.759	0.740	0.647
100	0.989	0.989	0.987	0.989	0.979	0.901	0.899
200	1.000	1.000	1.000	1.000	1.000	0.990	0.989
1000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
			$\rho_{xy}=0.8 \quad \rho_x=0.5 \quad \rho_y=0.9$				
50	0.527	0.537	0.506	0.512	0.437	0.405	0.414
100	0.864	0.864	0.852	0.864	0.809	0.668	0.676
200	0.992	0.994	0.990	0.992	0.991	0.914	0.911
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			$\rho_{xy}=0.8 \quad \rho_x=0.8 \quad \rho_y=0.5$				
50	0.848	0.846	0.807	0.848	0.800	0.735	0.637
100	0.991	0.991	0.988	0.991	0.982	0.915	0.898
200	1.000	1.000	1.000	1.000	1.000	0.993	0.987
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2: Empirical Power Performance of the Tests (continued)

n	HM test	x^2 test	Logit-t	regresstion-t	t_{HAC}	$t_{\text{fixed}-b}$	t^*
			$\rho_{xy}=0.8 \quad \rho_x=0.8 \quad \rho_y=0.8$				
50	0.821	0.821	0.715	0.822	0.747	0.662	0.569
100	0.979	0.979	0.971	0.979	0.962	0.884	0.875
200	1.000	1.000	1.000	1.000	1.000	0.989	0.980
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			$\rho_{xy}=0.8 \quad \rho_x=0.8 \quad \rho_y=0.9$				
50	0.649	0.646	0.515	0.642	0.535	0.455	0.444
100	0.891	0.894	0.863	0.891	0.836	0.720	0.740
200	0.996	0.996	0.996	0.996	0.994	0.931	0.940
1000	1.000	1.000	1.000	1.000	1.000	0.962	1.000
			$\rho_{xy}=0.8 \quad \rho_x=0.9 \quad \rho_y=0.5$				
50	0.578	0.581	0.540	0.583	0.391	0.413	0.450
100	0.891	0.894	0.863	0.891	0.836	0.720	0.740
200	0.996	0.996	0.996	0.996	0.994	0.931	0.940
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			$\rho_{xy}=0.8 \quad \rho_x=0.9 \quad \rho_y=0.8$				
50	0.666	0.649	0.548	0.671	0.574	0.547	0.525
100	0.891	0.894	0.863	0.891	0.836	0.720	0.740
200	0.996	0.996	0.996	0.996	0.994	0.931	0.940
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			$\rho_{xy}=0.8 \quad \rho_x=0.9 \quad \rho_y=0.9$				
50	0.605	0.603	0.438	0.600	0.546	0.492	0.437
100	0.835	0.843	0.769	0.835	0.793	0.674	0.738
200	0.986	0.987	0.983	0.986	0.978	0.914	0.920
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Notes: The figures reported are the bootstrap powers for t^* and the size-adjusted powers for other tests. The DGP is described from Eq. (12) to Eq. (16). The HM test is the market timing test in Henricksson and Merton (1981). The x^2 test denotes the independence test. Logit-t and regression-t represent the conventional t-test in the logit model and regression model, respectively. t_{HAC}, $t_{\text{fixed}-b}$, and t^* denote the HAC robust t-test with the standard normal distribution, fixed-b asymptotics, and empirical distribution from the naive block bootstrap, respectively. Refer to Section 2 for further details. The figures reported are the rejection frequencies at the 5% nominal significance level, based on 5,000 replications for the asymptotic tests, and 1,000 for t^* with 1,000 re-samples.

4 Empirical Illustrations

In this section, we present a simple empirical example to illustrate that the conventional tests for evaluating event forecast accuracy may be misleading when event forecasts and their corresponding realizations are serially correlated. We stress at the outset that the specification for a forecasting model is not the goal of this empirical study. Instead, we focus on evaluating the accuracy of event forecasts using the tests discussed in Section 2. Hence, an AR(1) forecasting model is sufficient to serve this purpose. Suppose that an AR(1) model $r_t = \alpha + \beta r_{t-1} + \varepsilon_t$ is used to forecast one-step-ahead market returns with a rolling scheme.[4] Let event forecasts Y_{t+1} and realizations X_{t+1} be defined as

$$Y_{t+1} = \begin{cases} 1 & \text{if } \hat{\alpha} + \hat{\beta} r_t > 0 \\ 0 & \text{otherwise} \end{cases},$$

and

$$X_{t+1} = \begin{cases} 1 & \text{if } r_{t+1} > 0 \\ 0 & \text{otherwise} \end{cases},$$

where $\hat{\alpha}$ and $\hat{\beta}$ denote LS estimates for α and β, respectively.

We collect monthly observations for five different indices over 1982:01—2007:02, including the All Ordinaries Index, Taiwan Weighted Index, Straits Times Index, Nikkei 225 Index, and TSX Composite Index. All the data are retrieved from Info Winner data bank. Let the price data be $\{p_t\}$. Then, the returns series is generated by $\{r_t\} = \ln P_t - \ln P_{t-1}$. The total sample of 301 return observations is split into two parts: the first 51 observations are used for estimating the parameters in AR(1) model, and then, the remaining 250 $(= n)$ are used to evaluate a post-sample one-step-ahead prediction. We choose $n = 250$ to avoid the low power of the tests.

The preliminary analysis for event forecasts and realizations is reported in Table 3. Obviously, the sample correlation coefficients indicate that Y_t and X_t are serially correlated for each index. The maintained IID assumption for the

[4]The rolling scheme, which discards the oldest observation when adding the latest one to estimate the parameters in the forecasting model, are more sensible than the fixed and recursive schemes, and is employed in our empirical study.

conventional tests may be violated. This can lead these tests to falsely reject the null hypothesis of no timing ability. In addition, except for the TSX Composite Index, the contemporary correlation between Y_t and X_t is so weak that the null hypothesis may not be rejected.

Table 3: Sample Correlation Coefficients

Index	(Y_t, Y_{t-1})	(X_t, X_{t-1})	(Y_t, X_t)
All Ordinaries	0.596	0.909	0.070
Taiwan Weighted	0.566	0.848	0.053
Straits Times	0.625	0.816	0.020
Nikkei 225	0.496	0.781	0.033
TSX Composite	0.599	0.864	0.611

The testing results are collated in Table 4. First, consider the All Ordinaries Index and Straits Times Index, the contemporary correlations of which are 0.07 and 0.02, respectively. Given the 5% significance level, the null is rejected by all the tests in the first group. On the contrary, all the tests that possess robust size in the second group do not reject the null. Similar results can be found except for the logit-*t* with a *p*-value of 0.305 in Nikkei 225 and with a *p*-value of 0.097 in Taiwan Weighted. The violation of the maintained IID assumption may be contributing to this contradiction. Although the null is rejected by all the tests for TSX Composite, the result must be interpreted with care. The rejection of the tests in the first group is more likely caused by their severe over-rejections. The testing result of t^*, on the other hand, can reflect the strong contemporary correlation between event forecasts and their corresponding realizations due to its robust size and reasonable power.

5 Conclusions

The conventional tests for assessing the accuracy of event forecasts, such as the HM test or regression *t*-test, rely on the maintained IID assumption on event forecasts and realizations. In practice, however, event forecasts obtained by truncating the point forecasts generated by estimated econometric models, may be serially correlated. Similarly, event realizations also have serial correlations. According to our simulation evidence, the violation of the maintained assumption contributes to

severe size distortions of these tests.

Table 4: Tests for the Accuracy of Event Forecasts

Index	HM test	x^2 test	Logit-*t*	regresstion-*t*	t_{HAC}	$t_{\text{fixed}-b}$	t^*
All Ordinaries	4.539	99.155	2.681	15.314	1.385	4.152	4.152
	(0.000)	(0.000)	(0.007)	(0.000)	(0.166)	[4.771]	(0.091)
Taiwan Weighted	11.598	113.574	1.659	12.165	0.789	1.952	1.952
	(0.000)	(0.000)	(0.097)	(0.000)	(0.430)	[4.771]	(0.324)
Straits Times	9.961	105.056	2.177	12.578	0.291	0.760	0.760
	(0.000)	(0.000)	(0.030)	(0.000)	(0.771)	[4.771]	(0.675)
Nikkei 225	16.743	118.336	1.027	9.408	0.561	1.180	1.180
	(0.000)	(0.000)	(0.305)	(0.000)	(0.575)	[4.771]	(0.528)
TSX Composite	7.968	103.792	3.162	15.982	2.219	5.857	5.857
	(0.000)	(0.000)	(0.000)	(0.000)	(0.026)	[4.771]	(0.023)

Notes: All the data are retrieved from Info Winner data bank. Monthly data is available over the period 1982:1－2007:2. All the tests are defined in Section 2. The values of the tests are reported in the first row for each index, and the p-values are in the parentheses. The value in [.] is the asymptotic critical value obtained from Table 1 in Kiefer and Vogelsang (2002) for the two-sided $t_{\text{fixed}-b}$ test at the 5% nominal significance level. The replications for the naive block bootstrap is 1,000 (= NB).

On the other hand, the fixed-*b* asymptotics for the HAC robust *t*-test offers a more accurate approximation than the standard normal distribution. Furthermore, the naive block bootstrap with an appropriately proper chosen block length can further improve the approximation to the sampling distribution of the HAC robust *t*-test. Our simulation evidence confirms these results. Therefore, the naive block bootstrap is strongly recommended for empirical applications. We also offer a simple empirical example to illustrate these tests. The testing results using naive block bootstrap are completely different from those with its conventional counterparts. The over-rejections of conventional tests can account for these empirical contradictions.

References

Andrew, D. W. K., (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817-858.

Andrew, D. W. K., (2002), "Higher-Order Improvements of a Computationally

Attractive K-Step Bootstrap for Extremum Estimators," *Econometrica*, 70, 119-162.

Breen, W., L. R. Glosten, and R. Jagannathan, (1989), "Economic Significance of Predictable Variations in Stock Index Returns," *The Journal of Finance*, 44, 1177-1189.

Davision, A. C. and P. Hall, (1993), "On Studentizing and Blocking Methods for Implementing the Bootstrap with Dependent Data," *Australian Journal of Statistics*, 35, 215-224.

Engel, C., (1994), "Can the Markov Switching Model Forecast Exchange Rates?" *Journal of International Economics*, 36, 151-165.

Gonçalves, S. and T. J. Vogelsang, (2006), "Block Bootstrap HAC Robust Tests: The Sophistication of the Naive Bootstrap," *Working Paper*.

Götze, F. and H. R. Künsch, (1996), "Second-Order Correctness of the Blockwise Bootstrap for Stationary Observations," *Annals of Statistics*, 24, 1914-1933.

Hall, P. and B. Y. Jing, (1996), "On Sample Reuse Methods for Dependent Data," *Journal of Royal Statistical Society Series B*, 58, 727-737.

Henriksson, R. D. and R. C. Merton, (1981), "On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills," *Journal of Business*, 54, 513-533.

Kiefer, N. M. and T. J. Vogelsang, (2002), "Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size," *Econometric Theory*, 18, 1350-1366.

Kiefer, N. M. and T. J. Vogelsang, (2005), "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory*, 21, 1130-1164.

Lahiri, S. N., (1996), "On Edgeworth Expansion and Moving Block Bootstrap for Studentized M-Estimators in Multiple Linear Regression Models," *Journal of Multivariate Analysis*, 56, 42-59.

Newey, W. K. and K. D. West, (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703-708.

Park, J. Y., (2003), "Bootstrap Unit Root Tests," *Econometrica*, 71, 1845-1895.

Politis, D. and H. White, (2004), "Automatic Block-Length Selection for the Dependent Bootstrap," *Econometric Reviews*, 23, 53-70.