Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 273

Improved Nonparametric Confidence Intervals in Time Series Regressions

Joseph P. Romano and Michael Wolf

February 2006

Improved Nonparametric Confidence Intervals in Time Series Regressions

Joseph P. Romano
Department of Statistics
Stanford University
Stanford, CA 94305
U.S.A.

Michael Wolf *
Institute for Empirical Research in Economics
University of Zurich
CH-8006 Zurich
Switzerland

February 2006

Abstract

Confidence intervals in econometric time series regressions suffer from notorious coverage problems. This is especially true when the dependence in the data is noticeable and sample sizes are small to moderate, as is often the case in empirical studies. This paper suggests using the studentized block bootstrap and discusses practical issues, such as the choice of the block size. A particular data-dependent method is proposed to automate the method. As a side note, it is pointed out that symmetric confidence intervals are preferred over equal-tailed ones, since they exhibit improved coverage accuracy. The improvements in small sample performance are supported by a simulation study.

JEL CLASSIFICATION NOS: C14, C15, C22, C32.

KEY WORDS: Bootstrap, Confidence Intervals, Studentization, Time Series Regressions, Prewhitening.

*We thank an anonymous referee for helpful comments that have led to an improved presentation of the paper.

1 Introduction

Regressions in macroeconomics and finance typically involve explanatory variables and/or error variables that exhibit serial dependence. It is well-known that standard regression theory does not apply to such settings. Appropriate asymptotic theory for time series regression has been developed and is routinely applied in practice; for example, see Hannan (1970) or White (2001). Over the last decade, however, the literature has shown that the finite sample properties of standard (or normal theory) methods are often lacking in practice; a common phenomenon is that confidence intervals undercover and that hypothesis tests overreject. We focus on the construction of confidence intervals in the remainder of the paper; but the ideas also apply to hypothesis tests.

A number of recent proposals have been made to construct confidence intervals that exhibit improved coverage accuracy. The two most common proposals, arguably, are prewhitening (Andrews and Mohanan, 1992; Newey and West, 1994) and block bootstrapping. To achieve asymptotic refinements based on the block bootstrap, it is important to use a studentization. In the general context of smooth functionals of sample means, this importance was pointed out by Davison and Hall (1993) and Götze and Künsch (1996). Lahiri (1996) established asymptotic refinements of an appropriate studentized bootstrap M-estimator, though he considered the case of fixed covariates. Hall and Horowitz (1996) and Andrews (2001) established asymptotic refinements of the block bootstrap for studentized statistics in the more general framework of GMM estimators, though they used somewhat restrictive dependence conditions.

Nevertheless, it seems that the (studentized) block bootstrap has not found wide approval of practitioners yet. As Horowitz (2001) puts it: "There are also unresolved problems in applying the bootstrap to a stationary, weakly dependent data generating process (DGP) when . . . a model that reduces the DGP to random sampling from a distribution is unavailable. The block bootstrap is the best-known method for implementing the bootstrap in such situations, but the performance of the block bootstrap in Monte Carlo experiments has been disappointing." We believe that a major contribution to this negative point of view has been the almost complete neglect in the relevant literature of the choice of the block size in practice. Since this choice can have a dramatic effect on the finite-sample properties, few practitioners should be willing to use the block bootstrap without any guidance as how to select the block size.

We discuss how to implement the studentized block bootstrap in the context of time series regressions. By detailing the approach and offering a concrete suggestion for the choice of the block size, we provide a useful method that is safe to apply in practice. In addition, it is mentioned that when two-sided confidence intervals are to be constructed, one should employ symmetric intervals as opposed to equal-tailed intervals, since the former tend to enjoy better coverage properties.

The remainder of the paper is organized as follows. Section 2 presents the model and the inference problem. Section 3 reviews the normal theory. Section 4 discusses the block bootstrap and details how to studentize in the context of time series regressions. Section 5 presents some asymptotic theory. Section 6 addresses the choice of the block size, which is an important model parameter. Section 7 considers finite-sample performance by means of a simulation study. Finally, Section 8 summarizes the findings. All tables appear in Section 9 at the end of the paper.

2 The Model

We consider the standard regression model

$$y_t = X_t'\beta + \epsilon_t, \; t = 1, \ldots, T$$

where $\beta \in \mathbb{R}^p$ is the unknown regression parameter and $(X_t', \epsilon_t)'$ is a weakly dependent, stationary sequence. The ordinary least squares (OLS) estimator for β is given by

$$\hat{\beta}_T = \left(\sum_{t=1}^{T} X_t X_t' \right)^{-1} \sum_{t=1}^{T} X_t y_t$$

A critical assumption to ensure its consistency is $E(X_t \epsilon_t) = 0, \; t = 1, \ldots, T$, that is, the regressors are uncorrelated with the error term. This assumption is typically implied by economic considerations, such as a rational expectation model. Under certain regularity conditions, the OLS estimator $\hat{\beta}_T$ will have an asymptotic normal distribution. We do not consider any nonstandard asymptotics, such as unit root regressions.

Interest focuses on constructing confidence intervals for a real-valued parameter $\theta = a'\beta$, where a is a fixed and known $p \times 1$ vector. Quite often θ will simply be a particular regression coefficient β_i of interest. The restriction to real-valued parameters is made mainly to give a natural setting for the construction of confidence intervals. But other scenarios could be considered as well. For example, one might be interested in testing a general linear constraint of the form $R\beta$, where R is a fixed and known $j \times p$ matrix. The method proposed can be easily adapted to testing problems; see Remark 4.1.

3 Normal Theory

Under nonrestrictive regularity conditions, $\sqrt{T}(\hat{\beta}_T - \beta) \stackrel{\mathcal{L}}{\Longrightarrow} N(0, \Sigma)$, where Σ is a positive-definite $p \times p$ matrix and $\stackrel{\mathcal{L}}{\Longrightarrow}$ denotes convergence in distribution (or convergence in law). This implies $\sqrt{T}(a'\hat{\beta}_T - \theta) \stackrel{\mathcal{L}}{\Longrightarrow} N(0, a'\Sigma a)$, which would allow one to construct an asymptotic normal theory confidence interval for θ if Σ were known. Unfortunately, Σ depends on the unknown underlying probability mechanism. The standard way of making inference is therefore to consistently estimate the limiting covariance matrix Σ by an estimator $\hat{\Sigma}_T$ and to pretend the distribution of $\sqrt{T}(a'\hat{\beta}_T - \theta)$ is given by $N(0, a'\hat{\Sigma}_T a)$, that is, to use the plug-in principle. As is well-known,

$$\Sigma = \lim_{T\to\infty} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T} E\epsilon_s X_s(\epsilon_t X_t)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}$$

Since the series $\{X_t\}$ is observed, consistent estimation of Σ only requires a consistent estimator of

$$J_T \equiv \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T} E\left[\epsilon_s X_s(\epsilon_t X_t)'\right]$$

The most popular approach to estimate J_T is by means of a kernel technique. In practice this involves choosing a real-valued kernel function $k(\cdot)$ and bandwidth S_T. The kernel $k(\cdot)$ typically satisfies the three conditions $k(0) = 1$, $k(\cdot)$ is continuous at 0, and $\lim_{x\to\pm\infty} k(x) = 0$. For more details on kernel estimation and a number of popular kernels, see Priestley (1981, Chapter 6) or Andrews (1991), among others. For related approaches and earlier references, see Robinson and Velasco (1997) and Den Haan and Levin (1997).

An important feature of a kernel is its characteristic exponent $1 \le q \le \infty$, determined by the smoothness of the kernel at the origin. Note that the bigger q, the smaller is the asymptotic bias of a kernel variance estimator; on the other hand, only kernels with $q \le 2$ yield estimates that are guaranteed to be positive semi-definite in finite samples. Most of the commonly used kernels have $q = 2$, such as the Parzen, Tukey-Hanning, and Quadratic-Spectral (QS) kernels, but exceptions do exist. For example, the Bartlett kernel has $q = 1$ and the Truncated kernel has $q = \infty$. For a broader discussion on this issue, see Priestley (1981, Chapter 6) for example.

Once a particular kernel $k(\cdot)$ has been chosen for application, one must pick the bandwidth S_T. Several automatic methods, based on various asymptotic optimality criteria, are available to this end; for example, see Andrews (1991) and Newey and West (1994). Note that the 'optimal' bandwidth generally depends on the underlying stochastic mechanism generating the data, the choice of the kernel $k(\cdot)$, and the sample size T.

Denote the kernel estimator of J_T by $\hat{J}_T$. The kernel estimator of Σ is then defined as

$$\hat{\Sigma}_T = \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \hat{J}_T \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}$$

Finally, the kernel estimator of the limiting variance $a'\Sigma a$ of $\hat{\theta}_T$ is given by

$$\hat{\sigma}_T^2 \equiv a'\hat{\Sigma}_T a \tag{1}$$

and one can construct normal theory confidence intervals based on this estimator.

Unfortunately, normal theory intervals often work unsatisfactorily in small samples, especially when the dependence structure of the underlying data is strong; again, see Andrews

(1991) and Newey and West (1994). The method of prewhitening, dating back to Press and Tukey (1956), has been suggested. For details the reader is referred to Priestley (1981, Chapter 7) and Andrews and Monahan (1992), among others. According to empirical studies in Andrews and Monahan (1992) and Newey and West (1994), confidence intervals based on prewhitened kernel estimators indeed exhibit improved finite sample performance, though they are still not perfectly satisfactory.

4 The Block Bootstrap and How to Studentize

Given the success of the bootstrap in regression settings with independent observations (Wu, 1986), it is natural to apply an appropriate bootstrap method to time series regressions. Such a method has to take into account the time series structure of both the regressors and the error variables. In the absence of semi-parametric structural models for these variables, such as ARMA or VAR models with i.i.d. innovations, the common approach is to resample blocks of data. For ease of notation, let $Z_t = (X_t', y_t)'$, so the observed data is $\{Z_1, \ldots, Z_T\}$. Also denote the true probability mechanism by P. The most popular time series bootstrap is the block bootstrap due to Künsch (1989) and Liu and Singh (1992). It considers overlapping blocks of size b, namely $Y_t = \{Z_t, \ldots, Z_{t+b-1}\}$, $t = 1, \ldots, T - b + 1$. Assuming for the moment $T = lb$, the method selects l blocks Y_t^* at random and with replacement from the $T - b + 1$ available blocks and concatenates them to arrive at the pseudo sequence $\{Y_1^*, \ldots, Y_l^*\} = \{Z_1^*, \ldots, Z_T^*\}$; in case T is not a multiple of b, one would do the same with the smallest l such that $T < lb$ and then truncate the pseudo sequence at T observations. Denote by P_T^* the bootstrap distribution (conditional on the observed data) of the pseudo sequence $\{Z_1^*, \ldots, Z_T^*\}$. Let $\hat{\beta}_T^*$ be the OLS estimator of β computed from the pseudo sequence and let $\hat{\theta}_T^*$ be the corresponding linear combination $a'\hat{\beta}_T^*$. A straightforward bootstrap approximation of the sampling distribution of the OLS estimator $\hat{\beta}_T$ is then

$$\mathcal{L}_P\{\hat{\theta}_T - \theta\} \approx \mathcal{L}_{P_T^*}\{\hat{\theta}_T^* - \hat{\theta}_T\} \tag{2}$$

Here, the general notation $\mathcal{L}_Q\{W\}$ denotes the law of a random variable W under a probability mechanism Q.

The relation (2) can now be used to construct an approximate confidence interval for θ. This particular bootstrap approximation is usually referred to as the hybrid bootstrap (Hall, 1992) or the basic bootstrap (Davison and Hinkley, 1997); we shall use the latter notation henceforth. A problem with the block bootstrap is that, due to ‘edge effects’, $\hat{\theta}_T$ is not equal to $\theta(P_T^*)$, the parameter θ corresponding to the bootstrap distribution P_T^*. Lahiri (1992), in the context of the sample mean, showed that this failure of the block bootstrap has negative second-order effects that can be remedied by employing the parameter of the bootstrap distribution, $\theta(P_T^*)$, in the centering. In our application this would correspond to the approximation

$$\mathcal{L}_P\{\hat{\theta}_T - \theta\} \approx \mathcal{L}_{P_T^*}\{\hat{\theta}_T^* - \theta(P_T^*)\}$$

It might be useful to further clarify what this parameter is exactly. (We thank a referee for this suggestion and corresponding details.) Focusing on β first, we have

$$\beta = \arg\min E\left[\frac{1}{T}\sum_{t=1}^{T}(y_t - X_t'\beta)^2\right]$$

and therefore

$$\beta = \left[E\left(\frac{1}{T}\sum_{t=1}^{T}(X_tX_t')\right)\right]^{-1} E\left(\frac{1}{T}\sum_{t=1}^{T}X_ty_t\right)$$

Because $\beta(P_T^*)$ is the bootstrap analog of β it follows that

$$\beta(P_T^*) = \left[E^*\left(\frac{1}{T}\sum_{t=1}^{T}(X_t^*X_t^{*\prime})\right)\right]^{-1} E^*\left(\frac{1}{T}\sum_{t=1}^{T}X_t^*y_t^*\right)$$

where E^* denotes the expectation with respect to the bootstrap distribution. Defining weights $\alpha_T(t) = (T-b+1)^{-1}\min(t/b, 1, (T-b+1)/b)$, the 'edge effects' of the block bootstrap now imply that

$$\beta(P_T^*) = \left[\sum_{t=1}^{T}\alpha_T(t)X_tX_t'\right]^{-1}\sum_{t=1}^{T}\alpha_T(t)X_ty_t$$

which is not equal to $\hat{\beta}_T$. Finally, $\theta(P_T^*) = a'\beta(P_T^*)$, which is not equal to $\hat{\theta}_T = a'\hat{\beta}_T$.

Andrews (2002) has suggested an alternative approach to recentering the bootstrap distribution; see also Hall and Horowitz (1996) and Lahiri (1996). It consists of changing the way one computes the bootstrap estimator of β, but then centering it by subtracting $\hat{\beta}_T$ again. The alternative bootstrap estimator is defined as

$$\bar{\beta}_T^* = \arg\min \frac{1}{T}\sum_{t=1}^{T}\left[(y_t^* - X_t^{*\prime}\beta)^2 - E^*\left(g_t^*\left(\hat{\beta}_T\right)\right)'\beta\right]$$

where $g_t^*\left(\hat{\beta}_T\right)$ is the score for observation t with bootstrap data and evaluated at $\hat{\beta}_T$; in particular, here $g_t^*\left(\hat{\beta}_T\right) = -X_t^*\left(y_t^* - X_t^{*\prime}\hat{\beta}_T\right)$. For the block bootstrap

$$E^*\left(\frac{1}{T}\sum_{t=1}^{T}g_t^*\left(\hat{\beta}_t\right)\right) = \sum_{t=1}^{T}\alpha_T(t)g_t\left(\hat{\beta}_T\right)$$

which in general is not zero and therefore $\bar{\beta}_T^*$ is in general not equal to $\hat{\beta}_T^*$. In this approach, one should use the approximation

$$\mathcal{L}_P\{\hat{\theta}_T - \theta\} \approx \mathcal{L}_{P_T^*}\{\bar{\theta}_T^* - \hat{\theta}_T\}$$

where $\bar{\theta}_T^* = a'\bar{\beta}_T^*$. This second approach seems to have been favored by the time series literature that considers more complicated settings than just smooth functions of means, such as the OLS estimator.

But the simplest solution avoids any extra computations at all. Politis and Romano (1992) introduced the circular block bootstrap where the original data are 'wrapped' in a circle prior to resampling blocks; edge effects are thus eliminated. This third approach ensures that $\hat{\theta}_T = \theta(P_T^*)$ for linear statistics $\hat{\theta}_T$, such as OLS estimators and linear combinations of them. (It would not work in more complicated settings, though.)

Fitzenberger (1997) applied the basic block bootstrap to time series regressions. However, its finite sample performance is not superior to normal theory intervals. The reason for this 'disappointment' has its roots in the much-studied, simpler setting of the sample mean for i.i.d. observations. It is well-known (e.g., Hall, 1992) that in this setting the basic bootstrap does not provide an asymptotic refinement over the CLT normal interval in the sense that both are only first order correct. To achieve second order correctness, a more sophisticated method such as the studentized bootstrap or the BC_a bootstrap has to be employed. This result carries over to the dependent case. Davison and Hall (1993) and Götze and Künsch (1996), abbreviated by GK in the remainder of this paper, considered inference for smooth functions of means in the context of stationary, dependent observations. They showed that the basic bootstrap is only first order correct while the studentized bootstrap provides an asymptotic refinement, at least under regularity conditions that ensure an Edgeworth expansion. It turns out to be important that the studentization be done in a certain way; see the discussion below. Moreover, the block bootstrap distribution needs to be centered around the mean of the bootstrap distribution rather than the sample mean, a problem that could be avoided by the use of the circular bootstrap again. Since the OLS estimator can be expressed as a smooth function of appropriate sample means, the corresponding theory actually follows from the work of GK; see Section 5.

The studentized bootstrap, together with the proper centering, leads to the approximation

$$\mathcal{L}_P\{(\hat{\theta}_T - \theta)/\hat{\sigma}_T\} \approx \mathcal{L}_{P_T^*}\{(\hat{\theta}_T^* - \theta(P_T^*))/\hat{\sigma}_T^*\} \tag{3}$$

which again can be used to construct a confidence interval for θ. Here, $\hat{\sigma}_T$ is an estimator of the limiting standard deviation of $\hat{\theta}_T$ and $\hat{\sigma}_T^*$ is an estimator of the limiting standard deviation of $\hat{\theta}_T^*$.

Following GK, these two estimators do not have the same functional form. Since the bootstrap sequence is generated by concatenating i.i.d. blocks of data, one can exploit this particular dependence structure to arrive at the following 'natural' estimator $\hat{\sigma}_T^*$ in the context of time series regressions: Assuming for simplicity that $T = lb$, where b is the block size used to construct the block bootstrap sequence, define

$$\hat{V}_t^* = X_t^*(y_t^* - X_t^* \hat{\beta}_T^*), \quad t = 1, \ldots, T$$

$$\zeta_j = \frac{1}{\sqrt{b}} \sum_{t=1}^{b} \hat{V}_{(j-1)b+t}^*, \quad j = 1, \ldots, l$$

and

$$\hat{\Sigma}_T^* = \left(\frac{1}{T} \sum_{t=1}^{T} X_t^* (X_t^*)' \right)^{-1} \frac{1}{l} \sum_{j=1}^{l} \zeta_j \zeta_j' \left(\frac{1}{T} \sum_{t=1}^{T} X_t^* (X_t^*)' \right)^{-1}$$

Then, the 'natural' estimator of the limiting standard deviation of $\hat{\theta}_T^*$ is given by

$$\hat{\sigma}_T^* = \sqrt{a' \hat{\Sigma}_T^* a}$$

On the other hand, the original sequence is a stationary time series without any 'special' dependence structure. It is therefore natural to use a kernel estimator for $\hat{\sigma}_T$ based on equation (1).

Remark 4.1 We have discussed how to use the studentized bootstrap in order to construct confidence intervals for a real-valued parameter $\theta = a'\beta$, where a is a fixed known $p \times 1$ vector. At times, interest might instead focus on a multivariate parameter $\theta = R\beta$, where R is a fixed and known $j \times p$ matrix. In this setting, it is more natural to consider hypothesis tests of the sort $H_0 : R\beta = r_0$. The test can be performed by approximating the sampling distribution of the Wald test statistic under the null hypothesis, using the bootstrap in the following way.

$$\mathcal{L}_P\{(R\hat{\beta}_T - r_0)'(R\hat{\Sigma}_T R')^{-1}(R\hat{\beta}_T - r_0)\} \approx_{H_0}$$
$$\mathcal{L}_{P_{T,0}^*}\{(R\hat{\beta}_T^* - r_0)'(R\hat{\Sigma}_T^* R')^{-1}(R\hat{\beta}_T^* - r_0)\}$$

When the bootstrap is used for the purposes of hypothesis testing, it is crucial that the bootstrap law $P_{T,0}^*$ satisfy the constraints of the null hypothesis (e.g., Politis et al., 1999, Section 1.8). For our application it has to be ensured that $R\beta(P_{T,0}^*) = r_0$. This cannot be achieved by simply resampling blocks of the observed data. One way of enforcing the null hypothesis in $P_{T,0}^*$, based on the circular block bootstrap, is the following. Denote by $\tilde{\beta}_T$ the constrained least squares estimators based on the observed data and satisfying $R\tilde{\beta}_T = r_0$. Also, let $\hat{\epsilon}_t = y_t - X_t'\hat{\beta}_T$ and $y_{t,0} = X_t'\tilde{\beta}_T + \hat{\epsilon}_t$. Then, $P_{T,0}^*$ resamples blocks from the 'null data' $(X_1', y_{1,0})', \ldots, (X_T', y_{T,0})'$. If the 'regular' block bootstrap is used instead, one needs to adjust for the edge effects in addition; see the discussion above.

5 Relevant Theory

5.1 Summary of the Theory Concerning Studentization

The OLS estimator $\hat{\theta}_T$ can be written as a smooth function of sample means. Hence, the first and second-order theory developed by GK applies directly; see their paper for a sufficient set of regularity conditions, such as moment and mixing conditions. GK showed that in order to obtain asymptotic refinements over normal theory, it is crucial to use a low bias variance estimator in the studentization of the OLS estimator $\hat{\theta}_T$. In the computation of $\hat{\sigma}_T$ they consequently propose the Truncated kernel which, having characteristic exponent $q = \infty$, enjoys minimum asymptotic bias among all kernels.

With this choice, the bootstrap approximation (3) has error $O_P(n^{-3/4+\epsilon})$, where ϵ is a small number (GK); in contrast, the approximation (2) has error larger than $O_P(n^{-1/2})$, as does normal theory. It needs to be pointed out, though, that the Truncated kernel may result in a negative variance estimate in finite sample. Should this occur, we propose to switch to a kernel estimator based on the QS kernel, which is guaranteed yield a nonnegative variance estimate as its characteristic exponent is $q = 2$. With this choice, the bootstrap approximation (3) has error $O_P(n^{-2/3+\epsilon})$, where ϵ is a small number (GK).

5.2 Some Theory Concerning Symmetrization

The improvement in the approximation of the sampling distribution of the (studentized) OLS estimator due to the bootstrap results in enhanced coverage accuracy of one-sided confidence intervals compared to normal theory (GK). One the other hand, two-sided equal-tailed bootstrap confidence intervals are not more accurate (in the sense of the rate of convergence to the nominal confidence level) than two-sided normal theory confidence intervals. The same results hold true in the simpler setting of the sample mean with i.i.d. data (Hall, 1992). Still, in this setting, equal-tailed bootstrap confidence intervals provide an improvement in terms of the constant (Hall, 1992).

When two-sided confidence intervals are desired, an alternative to equal-tailed intervals are symmetric intervals, based on the bootstrap approximation of the two-sided sampling distribution function

$$\mathcal{L}_P\{|\hat{\theta}_T - \theta|/\hat{\sigma}(\hat{\theta}_T)\} \approx \mathcal{L}_{P_T^*}\{|\hat{\theta}_T^* - \theta(P_T^*)|/\hat{\sigma}^*(\hat{\theta}_T^*)\}$$

In many contexts, such intervals are more accurate than two-sided normal theory intervals. Examples include the sample mean with i.i.d. data and OLS estimators for regression parameters with independent data (Hall, 1992). More generally, Hall and Horowitz (1996) and Andrews (2001) have obtained such results in the context of GMM estimators with stationary, dependent data, covering the case of OLS estimators. However, for this they have to assume a stricter dependence condition: the asymptotic variances of the estimators of interest can only depend on a finite (and known) number of correlations, so that kernel variance estimators are dispensed with.

It should be mentioned that Lahiri (1996) studied asymptotic refinements of the studentized block bootstrap for general M-estimators in multiple regressions, which include the OLS estimator as a special case. But he considered a setting where the covariates X_t are fixed and known while the error terms ϵ_t are a stationary, dependent sequence. This setting would be inappropriate for most economic applications.

Therefore, the results in the literature so far about asymptotic refinements of symmetric block bootstrap confidence intervals do not cover the setting of this paper. We now present some arguments to fill this gap.

Let $J_T(x, P) = P\{(\hat{\theta}_T - \theta)/\hat{\sigma}_T \le x\}$. We assume the existence of an Edgeworth expansion of the form

$$J_T(x, P) = \Phi(x) + n^{-\gamma} R_1(x, P)\varphi(x) + n^{-\delta} R_2(x, P)\varphi(x) + o(n^{-\delta}) \tag{4}$$

Here $\delta > \gamma$, $R_1(x, P)$ is an even, differentiable function of x, $R_2(x, P)$ is an odd function of x (or at least not even, for our purposes), Φ denotes the standard normal c.d.f. with density φ. For the existence of such expansions, see Hall (1992) and Lahiri (2003). Sufficient conditions are often stated in terms of smoothness and moment assumptions, and the expansion usually holds uniformly in x. In particular, in both the independent and dependent cases, γ is typically equal to 0.5; see Subsection 6.4.4. of Lahiri (2003). The value of δ is usually 1 in i.i.d. problems, but is typically < 1 in dependent cases, and its exact value depends heavily on the estimator $\hat{\sigma}_T$. Our heuristic considerations below do not require knowing the exact value, but we will henceforth assume $\delta \le 2\gamma \le 1$. We would like to argue that the coverage error of equal-tailed intervals is larger than that of symmetric intervals. The following heuristics generalize those in Hall (1992).

The expansion (4) implies a Cornish-Fisher type expansion for quantiles as follows. Let $q_T(1-\alpha, P)$ denote a $1-\alpha$ quantile of J_T. Then, it easily follows that, for some function R_3 (depending on R_1 and R_2),

$$q_T(1-\alpha, P) = z_{1-\alpha} - n^{-\gamma} R_1(z_{1-\alpha}, P) + n^{-\delta} R_3(z_{1-\alpha}, P) + o(n^{-\delta})$$

If these quantile are known, a theoretical equal-tailed confidence interval of nominal level $1-\alpha$ is given by

$$\{\tilde{\theta} : \; q_T(\alpha/2, P) \le (\hat{\theta}_T - \tilde{\theta})/\hat{\sigma}_T \le q_T(1-\alpha/2, P)\}$$

The bootstrap in essence replaces $q_T(\cdot, P)$ by $\hat{q}_T(\cdot)$ satisfying an expansion of the form

$$\begin{aligned} \hat{q}_T(1-\alpha) &= z_{1-\alpha} - n^{-\gamma} \hat{R}_1 - n^{-\delta} \hat{R}_3 + o_P(n^{-\delta}) \\ &= z_{1-\alpha} - n^{-\gamma} R_1(z_{1-\alpha}, P) + O_P(n^{-\gamma} |\hat{R}_1 - R_1(z_{1-\alpha}, P)|) + O_P(n^{-\delta}) \end{aligned}$$

By plugging both $\hat{q}_T(1-\alpha/2)$ into (4) and subtracting the expression when plugging in $\hat{q}_T(\alpha/2)$, the evenness of the function R_1 implies that the expansion for the resulting coverage error does not have a term of order $n^{-\gamma}$, but the order of error is at best of order $n^{-\delta}$. However, the coverage error is at best $O(n^{-\delta})$ (and possibly larger, depending on $O_P(|\hat{R}_1 - R_1|)$).

We will now argue that coverage of symmetric bootstrap confidence intervals is $o(n^{-\delta})$ (and is usually much smaller depending on higher order Edgeworth approximations). To appreciate why, (4) implies an expansion for

$$J_{|\cdot|,T}(x, P) = P\{|\hat{\theta}_T - \theta|/\hat{\sigma}_T \le x\}$$

(by evaluating (4) at both x and subtracting it at $-x$). Using the evenness of R_1 and φ, we find

$$J_{|\cdot|,T}(x,P) = \Phi(x) - \Phi(-x) + n^{-\delta}R_4(x,P)\varphi(x) + o(n^{-\delta}) \tag{5}$$

where $R_4(x,P) = R_2(x,P) - R_2(-x,P)$. If $s_T(1-\alpha,P)$ denotes a $1-\alpha$ quantile of $J_{|\cdot|,T}(\cdot,P)$, then (5) implies

$$s_T(1-\alpha,P) = z_{1-\alpha/2} - n^{-\delta}0.5R_4(z_{1-\alpha/2},P) + o(n^{-\delta})$$

A theoretical symmetric confidence interval of nominal level $1-\alpha$ is given by

$$\{\tilde{\theta}: \ |\hat{\theta}_T - \tilde{\theta}|/\hat{\sigma}_T \leq s_T(1-\alpha,P)\}$$

The bootstrap replaces $s_T(\cdot,P)$ by an estimate $\hat{s}_T(\cdot)$ which should satisfy

$$\hat{s}_T(1-\alpha) = z_{1-\alpha/2} - n^{-\delta}0.5\hat{R}_4 + o_P(n^{-\delta}) \tag{6}$$

which equals

$$z_{1-\alpha/2} - n^{-\delta}0.5R_4(z_{1-\alpha/2},P) + o_P(n^{-\delta}) \tag{7}$$

if $|\hat{R}_4 - R_n| = o_P(1)$. To evaluate the coverage error, plug (6) into (7) and use the delta method to deduce the coverage error of the symmetric bootstrap confidence interval is of order $o(n^{-\delta})$.

6 Choice of the Block Size

The application of the block bootstrap requires a choice of the block size b. Asymptotic theory typically only requires that $b \to \infty$ and that $b/T \to 0$ as $T \to \infty$; for example, see Künsch (1989) and Politis and Romano (1992). But these requirements are of little practical help. The choice of the block size is a difficult but important problem, comparable to the choice of the bandwidth for kernel variance estimators. In the relevant literature this problem is quite often either ignored or delayed to future research, which is a regrettable state of affairs. Our aim is to propose an inference method for time series regressions that is not only of academic interest but will also find the approval of practitioners. Therefore, we feel the need to provide at least a reasonable ad hoc method that can be used in practice, though we are unable to completely solve this difficult problem (and it appears unlikely that a 'perfect' solution will ever be found).

A notable exception in the literature, dealing explicitly with the problem of choosing the block size, is Hall et al. (1996). They showed that the optimal block size (minimizing the asymptotic mean squared error or MSE) depends significantly on context and is given by $C(P)n^{1/k}$, where $C(P)$ is a constant and $k = 3$, 4, or 5 for the contexts of variance estimation, estimation of a one-sided distribution function, or estimation of a two-sided distribution function, respectively. The constant $C(P)$ depends on the underlying joint distribution P and the context but a way is suggested to estimate it in practice. The problem with trying to adopt their approach for our purposes is two-fold. First, all the asymptotic MSE calculations are done for the basic block bootstrap and thus would no longer be valid for the studentized block bootstrap. Second, for the estimation of a distribution function, $C(P)$ depends on the argument of that function, that is, on y in $F(y) = Prob_P\{Y \leq y\}$ for a general random variable Y. Since for the construction of a confidence interval one needs to estimate a quantile of a distribution, it seems that the corresponding y would first have to be found in some recursive fashion.

Instead, we will propose a method which can be applied to an arbitrary bootstrap method, whether studentized or not, and which immediately tackles the task of estimating a specific quantile as opposed to estimating the distribution function at a given point. To this end, we suggest to use a *calibration* method, a concept dating back to Loh (1987, 1988, 1991). One can think of the actual coverage level $1-\lambda$ of a block bootstrap confidence interval as a function of the block size b, conditional on the underlying probability mechanism P, the nominal confidence level $1-\alpha$, and the sample size T. The idea is now to adjust the 'input' b in order to obtain the actual coverage level close to the desired one. Hence, one can consider the block size calibration function $g : b \to 1-\lambda$. If $g(\cdot)$ were known, one could construct an 'optimal' confidence interval by finding $\tilde{b}$ that minimizes $|g(b) - (1-\alpha)|$ and use $\tilde{b}$ as the block size of the time series bootstrap; note that $|g(b) - (1-\alpha)| = 0$ may not always have a solution.

Of course, the function $g(\cdot)$ depends on the underlying probability mechanism P and is therefore unknown. We now propose a bootstrap method to estimate it. The idea is that in principle we could simulate $g(\cdot)$ if P were known by generating data of size T according to P and by computing confidence intervals for θ for a number of different block sizes b. This process is then repeated many times and for a given b one estimates $g(b)$ as the fraction of the corresponding intervals that contain the true parameter. The method we propose is identical except that P is replaced by an estimate $\hat{P}_T$.

Algorithm 6.1 (Choice of the Block Size)

1. Fit a model $\hat{P}_T$ to the observed data $(X_1', y_1)', \ldots, (X_T', y_T)'$.
2. Fix a selection of reasonable block sizes b.
3. Generate K pseudo sequences $((X_1^*)', y_1^*)_k' \ldots, ((X_T^*)', y_T^*)_k'$, $k = 1, \ldots, K$, according to $\hat{P}_T$. For each sequence, $k = 1, \ldots, K$, and for each b, compute a confidence interval $\mathrm{CI}_{k,b}$.
4. Compute $\hat{g}(b) = \#\{\theta(\hat{P}_T) \in \mathrm{CI}_{k,b}\}/K$.
5. Find the value of $\tilde{b}$ that minimizes $|\hat{g}(b) - (1-\alpha)|$.

The role of the semi-parametric model in Algorithm 6.1 can be compared to the role of the semi-parametric model in the prewhitening process for kernel variance estimation. Even if the model is misspecified, it should contain some information on the dependence structure of the true mechanism P that can be exploited to estimate $g(\cdot)$. In practice we suggest to employ a VAR model, whose order could be estimated by one of the well-known information criteria, say, in conjunction with bootstrapping the estimated residuals.

Remark 6.1 Note that Algorithm 6.1 is essentially a double bootstrap and therefore computationally more expensive, by an order of magnitude, than the application of the bootstrap method once the block size has been determined.

Remark 6.2 If a bootstrap method is used for hypothesis testing rather than confidence interval construction, an analogous algorithm can be used by focusing on the significance level of the test rather than on the confidence level of the interval. Note that in this case the semi-parametric model $\hat{P}_T$ should be replaced by a semi-parametric model $\hat{P}_{T,0}$ which satisfies the constraints of the the null hypothesis; the remaining details are straightforward.

7 Simulation Study

The purpose of this section is to compare the small sample performance of various methods to construct two-sided confidence intervals in time series regressions. Performance is measured in terms of estimated coverage probability of nominal 95% intervals, based on $M = 2,000$ repetitions. The methods included in the study are normal theory intervals as well as basic and studentized bootstrap intervals. A few words regarding the various methods are in order.

The normal theory intervals use the QS kernel, both for the standard interval and for the prewhitened interval. The prewhitening is done using a VAR(1) model. The automatic choice of bandwidth is the one of Andrews (1991). We also tried the one of Newey and West (1994) but the differences were not meaningful and so the corresponding results are not reported.

As was discussed in Section 5, one can hope to improve upon the equal-tailed basic bootstrap confidence intervals by both studentizing and symmetrizing. To judge the magnitude of the corresponding improvements, we include equal-tailed basic, equal-tailed symmetric, equal-tailed studentized, and symmetric studentized intervals in the study. The study uses the circular block bootstrap in order to avoid a recentering of the bootstrap distributions; see Section 4. The studentized bootstrap intervals use the Truncated kernel for the studentization of the statistic based on the original sample, where the bandwidth is equal to the block size of the block bootstrap (GK). In case the resulting estimator $\hat{\sigma}_T$ is negative, we switch to the QS kernel. (Depending on the data generating process, sample size, and block size used, for the choices detailed below, a negative estimate $\hat{\sigma}_T$ due to the Truncated kernel occurs with a frequency of up to 15%.) The following abbreviations are used to label the various confidence interval types.

- NT: normal theory interval
- NT-PW: prewhitened normal theory interval
- BA-ET: equal-tailed basic bootstrap interval
- BA-SYM: symmetric basic boostrap interval
- STUD-ET: equal-tailed studentized bootstrap interval
- STUD-SYM: symmetric studentized bootstrap interval

To generate the data, we consider the classic design of Andrews (1991), which has also been considered by Andrews and Monahan (1992), Fitzenberger (1997), and Politis et al. (1997), among others. However, to speed up the computations, we include only one non-constant regressor. The reason is that our data-dependent choice of the block size is computationally quite expensive. This not really a problem for an application to a (single) real data set. But it constitutes a problem in a simulation study when thousands of confidence intervals have to be computed.

The basic design therefore looks as follows

$$y_t = \beta_1 + \beta_2 x_t + \epsilon_t$$

Throughout, we are concerned with constructing confidence intervals for the regression parameter β_2. Without loss of generality, we set $\beta_1 = \beta_2 = 0$ when generating the data.

In the first model, AR(1)-HOMO, errors and regressors are independent AR(1) processes.

$$\text{AR(1)-HOMO:} \quad x_t = \rho x_{t-1} + \nu_{t,j} \text{ and } \epsilon_t = \rho \epsilon_{t-1} + \nu_t^\epsilon$$

Here, and for the following models, $\{\nu_t\}$ and $\{\nu_t^\epsilon\}$ are mutually independent white noise processes.

The second model, AR(1)-HET, is a variation of the first one in the sense that multiplicative heteroskedasticity is overlaid on the errors.

$$\text{AR(1)-HET:} \quad x_t = \rho x_{t-1} + \nu_{t,j}\ ,\ \tilde{\epsilon}_t = \rho \tilde{\epsilon}_{t-1} + \nu_t^\epsilon \text{ and } \epsilon_t = |x_t|\, \tilde{\epsilon}_t$$

In the third model, MA(1)-HOMO, both the errors and the regressors are independent MA(1) processes.

$$\text{MA(1)-HOMO:} \quad x_t = \nu_{t,j} + \xi \nu_{t-1} \text{ and } \epsilon_t = \nu_t^\epsilon + \theta \nu_{t-1}^\epsilon$$

For all three models, $\{\nu_t\}$ and $\{\nu_t^\epsilon\}$ are independent i.i.d. innovation sequences, having a standard normal distribution. The values considered for the parameters ρ and ξ are 0.2, 0.5, and 0.8, respectively. The sample size is $T = 64$. All bootstrap methods use the data-dependent choice of block size of Algorithm 6.1. The input block sizes for the algorithm are $b = 5, 12, 20$. The semi-parametric model in Algorithm 6.1 is a VAR(1) with bootstrapping the fitted residuals. For the latter, the circular block bootstrap with a small block size ($b = 5$) is used to capture some left-over dependence in the residuals when the VAR(1) is misspecified, as is the case for the MA(1)-HOMO model. Tables 1–3 present the results. They can be summarized as follows.

- In accordance with Davison and Hall (1993) and Götze and Künsch (1996), the basic bootstrap does not improve upon normal theory.
- In accordance with Andrews and Monahan (1992) and Newey and West (1994), prewhitening is useful in normal theory intervals.
- In accordance with Davison and Hall (1993) and Götze and Künsch (1996), the studentized bootstrap improves upon the basic bootstrap and normal theory.
- In accordance with Subsection 5.2, symmetric bootstrap confidence intervals generally improve upon equal-tailed ones, although this effect is noticeably stronger for the basic bootstrap compared to the studentized bootstrap.

We are also interested in a formal test whether the method we suggest, STUD-SYM, is superior to the other methods. To this end, for each scenario and each other competitor, we compute a p-value based on a permutation test. The details are as follows. Let $1 - \alpha$ be the nominal coverage level and let $1 - \lambda_1$ and $1 - \lambda_2$ be the two actual coverage levels, where we use the subscript 1 for the competing method and the subscript 2 for STUD-SYM. Further, define the two absolute coverage error as $\eta_1 = |(1 - \lambda_1) - (1 - \alpha)| = |\alpha - \lambda_1|$ $\eta_2 = |\alpha - \lambda_2|$, respectively. We want to test

$$H_0 : \eta_1 - \eta_2 = 0 \quad \text{vs.} \quad H_1 : \eta_1 - \eta_2 > 0 \tag{8}$$

The simulations yield the estimated coverage probabilities for the two methods, denoted by $1 - \hat{\lambda}_1$ and $1 - \hat{\lambda}_2$. The observed test statistic is then given as

$$TS_{obs} = \hat{\eta}_1 - \hat{\eta}_2 = |\alpha - \hat{\lambda}_1| - |\alpha - \hat{\lambda}_2|$$

The corresponding p-value is obtained via a suitable permutation test. To this end we consider the indicator pairs $(I_{1,m}, I_{2,m})$ for $m = 1, \ldots, 2,000$. Here $I_{2,m}$ equals 1 if the STUD-SYM confidence interval computed from the mth simulated data set contains the true parameter $\beta_2 = 0$, and it equals zero otherwise. $I_{1,m}$ is defined analogously for the competing method. The idea now is to randomly permute the ordering within each indicator pair. In this way, we derive the distribution of the test statistic under the null hypothesis $H_0 : \eta_1 - \eta_2 = 0$. The exact null distribution is computed from all $2^{2,000}$ possible permutations and is therefore not feasible. So we resort to a stochastic approximation based on $K = 9,999$ permutations chosen at random; see Romano (1989). From each permuted set of indicator pairs, we compute the test statistic, resulting in $TS_{0,k}$ for $k = 1, \ldots, K$. Following (4.11) in Davison and Hinkley (1997), the p-value for the testing problem (8) is now obtained as

$$\hat{p} = \frac{\#\{TS_{0,k} \geq TS_{obs}\} + 1}{K + 1} \tag{9}$$

The results are contained in the last column of Tables 1–3, respectively. In most scenarios, the outperformance of STUD-SYM over the competitors is significant with the exception of STUD-ET. The latter method performs six times worse than STUD-SYM (out of which two times significantly) and tree times better than STUD-SYM (out of which never significantly).

Since we consider nine different scenarios (three models with three different parameter values each), there are nine p-values corresponding to the comparison of any given confidence interval with STUD-SYM. It is then also of interest to combine these individual p-values into a test of 'overall' outperformance. To this end we employ the Z-transform test dating back to Stouffer et al. (1949, page 49); also see Mosteller and Bush (1954). The test works as follows. Let $\hat{p}_1, \ldots, \hat{p}_9$ denote the nine individual p-values, let $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal distribution, and let $\Phi^{-1}(\cdot)$ denote its inverse. Then define the combined test statistic as

$$Z_S = \frac{\sum_{i=1}^{9} \Phi^{-1}(\hat{p}_i)}{\sqrt{9}} = \frac{\sum_{i=1}^{9} Z_i}{3}$$

and the resulting p-value as

$$\hat{p}_S = \Phi(Z_s)$$

The intuition is that if all nine individual null hypotheses are true, then the Z_i behave like independent standard normal random variates and, as a result, so does Z_S. The results are contained in Table 4. It is shown that, in an overall sense, STUD-SYM outperforms all its competitors significantly, including STUD-ET.

Alternatively, we carry out a 'direct' combined permutation test. To this end let $\eta_{2,i}$ be the absolute coverage error of STUD-SYM in the ith scenario, $i = 1, \ldots, 9$. Then the overall absolute coverage error of STUD-SYM is defined as

$$\tilde{\eta}_2 = \sum_{i=1}^{9} \eta_{2,i}$$

The overall absolute coverage error of the competing method, $\tilde{\eta}_1$ is defined analogously. We want to test

$$H_0 : \tilde{\eta}_1 - \tilde{\eta}_2 = 0 \quad \text{vs.} \quad H_1 : \tilde{\eta}_1 - \tilde{\eta}_2 > 0 \tag{10}$$

The p-value is obtained via a permutation method similarly to (9), also based on a random set of $R = 9,999$ permutations. The results are contained in Table 5. Again, it is shown that, in an overall sense, STUD-SYM outperforms all its competitors significantly, including STUD-ET.

8 Conclusions

In this paper, the use of the studentized block bootstrap for time series regressions was proposed. The relevant second-order theory follows from previous work. On the other hand, the important problem of the choice of the block size had been rather neglected so far. We have offered a practical suggestion to deal with this problem. Its main disadvantage, the computational cost, will diminish over time. The finite-sample performance of various confidence interval types, measured by the empirical coverage probability, was examined via a simulation study. Based on the results of this study, the studentized block bootstrap does indeed yield improved performance compared to normal theory intervals and also the basic bootstrap. Furthermore, two-sided symmetric bootstrap confidence intervals are superior to two-sided equal-tailed bootstrap intervals.

References

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

Andrews, D.W.K. (2002). Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators. *Econometrica* **70**, 119–162.

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

Davison, A.C and Hall, P. (1993). On studentizing and blocking methods for implementing the bootstrap with dependent data. *Australian Journal of Statistics* **35**, 215–224.

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Fitzenberger, B. (1997). The moving blocks bootstrap and robust inference for linear least squares and quantile regressions. *Journal of Econometrics* **82**, 235–287.

Götze, F. and Künsch, H.R. (1996). Second order correctness of the blockwise bootstrap for stationary observations. *Annals of Statistics* **24**, 1914–1933.

Den Haan, W.J. and Levin, A. (1997). A practitioner's guide to robust covariance estimation. In *Handbook of Statistics, Robust Inference, Vol. 15*, 299–342. Edited by Maddala, G.S. and Rao, C.R., Elsevier Science B.V..

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

Hall, P. and Horowitz, J. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica* **64**, 891–916.

Hall, P., Horowitz J.L., and Jing, B.-Y. (1996). On blocking rules for the bootstrap with dependent data. *Biometrika* **50**, 561–574.

Hannan, E.J. (1970). *Multiple Time Series*. John Wiley, New York.

Horowitz, J. (2001). The bootstrap and hypothesis tests in econometrics. *Journal of Econometrics* **100**, 37–40.

Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* **17**, 1217–1241.

Lahiri, S.N. (1992). Edgeworth correction by 'Moving block' bootstrap for stationary and nonstationary data. In *Exploring the Limits of Bootstrap*, 183–214. Edited by LePage, R. and Billard, L., John Wiley, New York.

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.

Lahiri, S.N. (1999). On second-order properties of the stationary bootstrap method for studentized statistics. In *Asymptotics, Nonparametrics, and and Time Series*, 683–712. Edited by Ghosh, S., Marcel Dekker, New York.

Lahiri, S.N. (2003). *Resampling Methods for Dependent Data.* Springer, New York.

Liu, R.Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap*, 225–248. Edited by LePage, R. and Billard, L., John Wiley, New York.

Loh, W.Y. (1987). Calibrating confidence coefficients. *Journal of the American Statistical Association* **82**, 155–162.

Loh, W.Y. (1988). Discussion of "Theoretical comparison of bootstrap confidence intervals" by P. Hall. *Annals of Statistics* **16**, 972–976.

Loh, W.Y. (1991). Bootstrap calibration for confidence interval construction and selection. *Statistica Sinica* **1**, 479–495.

Mosteller, F. and Bush, R.R. (1954) Selected quantitative techniques. In: *Handbook of Social Psychology*, Vol. 1 (G. Lindzey, ed.), pp 289–334. Addison-Wesley, Cambridgde, Massachusetts.

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

Politis, D.N. and Romano, J.P. (1992). A circular block-resampling procedure for stationary data. In *Exploring the Limits of Bootstrap*, 263–270. Edited by LePage, R. and Billard, L., John Wiley, New York.

Politis, D.N. and Romano, J.P. (1994). The stationary bootstrap. *Journal of the American Statistical Association* **89**, 1303–1313.

Politis, D.N., Romano, J.P., and Wolf, M. (1997). Subsampling for heteroskedastic time series. *Journal of Econometrics* **81**, 281–317.

Politis, D.N. Romano, J.P., and Wolf, M. (1999). *Subsampling.* Springer, New York.

Press, H. and Tukey, J. W. (1956). Power spectral methods of analysis and their application to problems in airplane dynamics. Bell Systems Monograph No. 2606.

Priestley, M.B. (1981). *Spectral Analysis and Time Series.* Academic Press, New York.

Robinson, P.M. and Velasco, C. (1997). Autcorrelation-robust inference. In *Handbook of Statistics, Vol. 15*, 267–298. Edited by Maddala, G.S. and Rao, C.R., Elsevier Science B.V..

Romano, J.P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Annals of Statistics* **17**, 141–159.

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., and Williams, R.M.Jr. (1949). *The American Soldier, Vol.1: Adjustment during Army Life.* Princeton University Press, Princeton.

White, H.L. (2001). *Asymptotic Theory for Econometricians*, revised edition. Academic Press, New York.

Wu, C.F. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics* **14**, 1261–1343.

9 Tables

Table 1: Estimated coverage probabilities of various confidence intervals with nominal level 95% and corresponding p-values. The first column describes the confidence interval; the second column contains the estimated coverage probabilities, based on $M = 2,000$ repetitions; and the third column contains the p-value for the null hypothesis that the listed confidence interval is equally good as STUD-SYM. Details are in Section 7.

AR(1)-HOMO model, $\rho = 0.2$		
Interval	Est. coverage	p-value
NT	93.4	9.1
NT-PW	93.2	4.8
BA-ET	91.8	0.0
BA-SYM	92.8	0.9
STUD-ET	93.1	0.9
STUD-SYM	94.3	
AR(1)-HOMO model, $\rho = 0.5$		
Interval	Est. coverage	p-value
NT	89.2	0.0
NT-PW	91.3	0.0
BA-ET	87.9	0.0
BA-SYM	89.7	0.0
STUD-ET	94.2	46.1
STUD-SYM	94.3	
AR(1)-HOMO model, $\rho = 0.8$		
Interval	Est. coverage	p-value
NT	82.2	0.0
NT-PW	89.8	0.0
BA-ET	81.5	0.0
BA-SYM	86.2	0.0
STUD-ET	94.2	8.4
STUD-SYM	95.2	

Table 2: Estimated coverage probabilities of various confidence intervals with nominal level 95% and corresponding p-values. The first column describes the confidence interval; the second column contains the estimated coverage probabilities, based on $M = 2,000$ repetitions; and the third column contains the p-value for the null hypothesis that the listed confidence interval is equally good as STUD-SYM. Details are in Section 7.

AR(1)-HET model, $\rho = 0.2$		
Interval	Est. coverage	p-value
NT	90.8	0.0
NT-PW	90.5	0.0
BA-ET	87.5	0.0
BA-SYM	88.9	0.0
STUD-ET	93.8	26.5
STUD-SYM	94.2	
AR(1)-HET model, $\rho = 0.5$		
Interval	Est. coverage	p-value
NT	87.4	0.0
NT-PW	89.2	0.0
BA-ET	84.0	0.0
BA-SYM	86.3	0.0
STUD-ET	94.2	30.3
STUD-SYM	94.5	
AR(1)-HET model, $\rho = 0.8$		
Interval	Est. coverage	p-value
NT	77.9	0.0
NT-PW	85.6	0.0
BA-ET	73.9	0.0
BA-SYM	80.8	0.0
STUD-ET	93.3	0.6
STUD-SYM	94.6	

Table 3: Estimated coverage probabilities of various confidence intervals with nominal level 95% and corresponding p-values. The first column describes the confidence interval; the second column contains the estimated coverage probabilities, based on $M = 2,000$ repetitions; and the third column contains the p-value for the null hypothesis that the listed confidence interval is equally good as STUD-SYM. Details are in Section 7.

MA(1)-HOMO model, $\xi = 0.2$		
Interval	Est. coverage	p-value
NT	93.4	37.2
NT-PW	93.1	21.1
BA-ET	91.7	0.2
BA-SYM	92.9	12.3
STUD-ET	93.9	76.2
STUD-SYM	93.6	
MA(1)-HOMO model, $\xi = 0.5$		
Interval	Est. coverage	p-value
NT	90.3	0.0
NT-PW	92.0	1.3
BA-ET	89.1	0.0
BA-SYM	89.8	0.0
STUD-ET	93.7	72.2
STUD-SYM	93.5	
MA(1)-HOMO model, $\xi = 0.8$		
Interval	Est. coverage	p-value
NT	92.0	0.0
NT-PW	93.9	16.0
BA-ET	91.2	0.0
BA-SYM	92.0	0.0
STUD-ET	94.7	70.5
STUD-SYM	94.5	

Table 4: Combined p-values for the null hypothesis that the listed confidence interval is as good or better than STUD-SYM. For each method, the combined p-value is obtained by applying the Z-transform test to the individual p-values from the nine different scenarios used in Tables 1–3. Details are in Section 7.

Interval	Combined p-value
NT	0.0
NT-PW	0.0
BA-ET	0.0
BA-SYM	0.0
STUD-ET	2.9
STUD-SYM	

Table 5: The p-values for the overall null hypothesis (10) that the listed confidence interval is as good or better than STUD-SYM. For each method, the p-value is obtained by applying a permutation test to the overall absolute coverage errors. Details are in Section 7.

Interval	Combined p-value
NT	0.0
NT-PW	0.0
BA-ET	0.0
BA-SYM	0.0
STUD-ET	1.3
STUD-SYM	