Non-nested Models and the Likelihood Ratio Statistic: A Comparison of Simulation and Bootstrap-based Tests

George Kapetanios and Melvyn Weeks

February 2003

DAE Working Paper No. 0308

Not to be quoted without permission

ABSTRACT

We consider an alternative use of simulation in the context of using the Likelihood-Ratio statistic to test non-nested models. To date simulation has been used to estimate the Kullback-Leibler measure of closeness between two densities, which in turn 'mean adjusts' the Likelihood-Ratio statistic. Given that this adjustment is still based upon asymptotic arguments, an alternative procedure is to utilise bootstrap procedures to construct the empirical density. To our knowledge this study represents the first comparison of the properties of bootstrap and simulation-based tests applied to non-nested tests. More specifically, the design of experiments allows us to comment on the relative performance of these two testing frameworks across models with varying degrees of nonlinearity. In this respect although the primary focus of the paper is upon the relative evaluation of simulation and bootstrap-based nonnested procedures in testing across a class of nonlinear threshold models, the inclusion of a similar analysis of the more standard linear/log-linear models provides a point of comparison.

Key Words and Phrases: Non-nested tests, simulation-based inference, bootstrap tests, nonlinear threshold models
JEL Classification: C15, C52

NON-NESTED MODELS AND THE LIKELIHOOD RATIO STATISTIC: A COMPARISON OF SIMULATION AND BOOTSTRAP-BASED TESTS

George Kapetanios[1] and Melvyn Weeks[2][3]
University of Cambridge

1 Introduction

For two non-nested models, say f and g, twice the likelihood ratio statistic $T_f = \bar{l}_f - \bar{l}_g$, where $\bar{l}_f$ and $\bar{l}_g$ are, respectively, the observed means of the log-likelihood and the pseudo log-likelihood, is not asymptotically distributed as a chi-square random variable. Following work by Pesaran and Pesaran (1993) and Weeks (1996), computational techniques have been used to affect adjustments to the test statistic in order to improve the finite sample size and power properties. However, this approach based upon the Modified Likelihood principle and due to the seminal work of Cox (1961), is still reliant upon a reference distribution which is valid asymptotically. In addition, as Orme (1994) attests, the existence of a large number of asymptotic equivalent variants of the Cox test statistic represents a formidable menu of choices for the applied econometrician. In the case of the numerator, various test statistics are based upon the use of different consistent estimators of the Kullback-Leibler (KL) measure of closeness. An additional set of variants of the Cox test statistic depend upon a number of asymptotically equivalent ways of estimating the variance of the test statistic.[4] In this context it is important to emphasise that simulation is used simply to provide an estimate of the test statistic. Nonetheless, Pesaran and

[1] Department of Economics, Queen Mary, University of London. Tel: 020 7882 5097 Email: g.kapetanios@qmul.ac.uk

[2] Faculty of Economics and Politics, University of Cambridge, Cambridge, CB3 9DD. Tel: 01223 335260 Email: mw217@econ.cam.ac.uk

[3] We would like to thank Hashem Pesaran and Les Godfrey for constructive comments.

[4] Orme (1994) presents a detailed analysis of these variants.

Pesaran (1993) note that *in principle it should be possible to use the simulation method both for the computation of the Cox statistic and for a derivation of a better small sample approximation to its distribution under the null* (p. 378).

An alternative approach based upon the seminal work of Efron (1979) with contributions by Hall (1986), Beran (1988), Hinkley (1988), and Coulibaly and Brorsen (1997), applies bootstrap-based procedures to evaluate the adequacy of nonnested models. In this context the focus is upon correcting the reference distribution rather than adjusting the test statistic and utilising limiting distribution arguments. This type of adjustment can, in a number of cases, be theoretically justified through Edgeworth expansions and under certain conditions result in improvements over classical asymptotic inference. The existence of a large menu of broadly equivalent test statistics is also relevant in the context of bootstrap-based inference. Surveys by Vinod (1993), Jeong and Maddala (1993), and Li and Maddala (1996), review a large number of variants including the double, recursive and weighted bootstrap. Hall (1998) notes that in many applications the precise nature of the bootstrap design is not stated.

In considering a number of bootstrap and asymptotic tests, we make the distinction between pivotal and non-pivotal test statistics. A test statistic is pivotal if it does not depend on unknown parameters; it is asymptotically pivotal if this condition holds only asymptotically. A number of authors, including Beran (1988), have demonstrated that the bootstrap distribution of an asymptotically pivotal test statistic provides a closer approximation to the true unknown distribution function than first-order classical asymptotics. This is in contrast to previous studies which have focussed on the substitution of a crude frequency simulator for the often intractable Kullback-Leibler (KL) measure of distance between two densities in the context of 'mean adjusting' the LR statistic. Again, for the purpose of our analysis, the distinction between working with a different reference distribution as opposed to alternative methods to construct the test statistic itself, is central.

In this paper we use a combination of simulation and bootstrap procedures to conduct nonnested hypothesis tests for linear and non-linear models. First, we test the linear versus loglinear regression model. The choice between linear and log-linear regression models continues to be an important issue for the applied econometrician. In general economic theory has little to offer and given that the two models are non-nested, classical inference founded upon the log-likelihood ratio statistic cannot be utilised. Following Beaudry and Koop (1993a) and Kapetanios (1998), we then use both simulation and bootstrap tests to differentiate between a class of nonnested threshold models, highlighting the performance of the tests in distinguishing different nonlinear mechanisms. Threshold models, introduced and discussed extensively by Tong (1978), Tong (1983), and Tong (1995) have been widely used as a framework for examining the presence of nonlinearity in empirical econometric models. The wide variety of classes of nonlinear threshold models necessitates the use of evaluation procedures to compare and discriminate between them. To our knowledge this study represents the first comparison of the properties of bootstrap and simulation-based tests applied to non-nested tests. More specifically, the design of experiments allows us to comment on the relative performance of these two testing frameworks across models with varying degrees of nonlinearity. In this respect although the primary focus of the paper is upon the relative evaluation of simulation and bootstrap-based nonnested procedures in testing across a class of nonlinear threshold models, the inclusion of a similar analysis of the more standard linear/log-linear models provides a point of comparison.

In the first set of experiments involving threshold models we examine whether nonnested hypotheses tests can distinguish between a simple nonlinear model such as that developed by Potter (1995) and more realistic multiregime models. In particular we examine the size and power properties of a test of a two versus three regime self exciting threshold autoregressive (SETAR) model. In the second setup the choice is between a SETAR trend and an EDTAR trend model where the linear parts of the models are the same.

It is important to emphasise that in the case of two and three regime SETAR models nonnested testing provides an alternative testing procedure when some parameters are not identified under the null hypothesis (see Davies (1977)). The main reason for utilising a nonnested testing framework in this case is the difficulty of carrying out the nested test. To see this note that the standard solution to the Davies problem requires modifications in this instance because even the threshold parameter which is in principle identified under the null hypothesis of a two-regime SETAR model is not known and needs to be estimated. We also note it is possible to consider model selection procedures to distinguish between the case of $m = 2$ and $m = 3$ or even to estimate the number of regimes, m. However, in motivating pairwise comparisons we argue that the distinction between $m = 2$ and $m = 3$ can be economically important. For example, the use of a 2-regime SETAR model in the modelling of a macroeconomic series driven by the business cycle, such as output or unemployment, has implications which are qualitatively distinct from the use of a 3-regime SETAR model or a SETAR models with $m > 3$. This is underlined in the literature on the asymmetry of the business cycle over recessions and expansions where a 2-regime characterisation of the business cycle, as in Potter (1995), is juxtaposed with a multi-regime characterisation as in Sichel (1994).

In utilising bootstrap tests our principal objective is to determine the extent to which the construction of the empirical distribution function of the test statistic represents an improvement over first-order asymptotic approximations. The extent to which the use of bootstrap procedures to approximate the sampling distribution of the likelihood ratio (LR) statistic enables a Bartlett-type adjustment to the asymptotics is central. In each case we utilise resampling to construct the empirical density of the likelihood ratio statistic, and consider a number of variants of bootstrap statistics. The first, a non-pivotal test statistic is the analog of the percentile method for the construction of confidence intervals. The other methods employ some form of standardisation to reduce the dependence of the empirical distribution function on unknown parameters..An alternative testing framework employ simulation-based methods to provide consistent location and scale adjustments to the LR statistic. In this instance we utilise simulation in the constriction of a number of alternative *test statistics,* whilst relying upon asymptotic arguments as a basis for inference.

The outline of the paper is as follows. In Section 2 we motivate the analysis by providing a brief overview of some key issues in the construction of bootstrap test statistics. In particular we examine the distinction between simulation and bootstrap error. In Section 3 we present the structure of the Cox test statistic and in doing so highlight the computational difficulties involved in the use of Cox's non-nested test. Section 4 presents results for testing linear and log-linear regression model utilising a number of asymptotic and bootstrap test statistics. In Section 5 we introduce a number of nonlinear threshold models and examine the finite sample performance of the same set of test statistics.

2 Bootstrap Statistics

The bootstrap distribution of a statistic can be defined as the exact finite sample distribution function evaluated at an estimate of the unknown parameters. As discussed by Singh (1981), Hall (1986), Hall (1992) and Brown (2000), bootstrapping a studentized statistic that is asymptotically pivotal will provide a closer approximation to the true distribution than the standard limiting distribution, with coverage differing from the nominal level by only $O_p(n^{-1})$ instead of $O_p(n^{-1/2})$, for independent observations. Hartigan (1986), Hall (1988) and Beran (1988) advocate the use of pivoting[5] as a device to reduce the error in rejection probability. Although much of the asymptotic theory for the bootstrap has been developed for the construction of confidence intervals, the well known duality between hypothesis testing and confidence intervals guarantees that any ranking of bootstrap variants for confidence intervals will hold in the case of hypothesis testing.

The drawback of this method has been noted by a number of authors including Tibshirani (1988) and more recently Horowitz (1995). The principal disadvantage is that studentizing requires an estimate of the standard deviation of the test statistic which in some cases can represent a poor approximation to the true value. Further, a pivoting procedure advocated by Beran (1988) requires the use of an inner bootstrap loop and as such there is an obvious trade-off between reduction in approximation error and the attendant computational burden. In addition, we note that the asymptotic theory is not informative in the absence of pivotalness. Since in most cases statistics are only asymptotically pivotal, then faced with a finite sample there is no theory-based ranking for pivotal versus non-pivotal bootstrap statistics.

[5] Note the difference between pivoting which refers to appropriately standardising a statistic to render it pivotal (or asymptotically pivotal) and prepivoting. Prepivoting has been suggested by Beran (1988) and involves bootstrapping the cumulative distribution function of a statistic rather than the statistic itself. Prepivoting has been shown to improve the asymptotic performance of the bootstrap on asymptotically pivotal statistics.

In a recent paper Kilian (1998) highlights the perception that following the seminal work of Singh (1981) and a comprehensive summary by Hall (1992) reviewing the asymptotic theory for bootstrap tests, the use of pivotal statistics can only reduce approximation error. Hall and Wilson (1991), Giersbergen and Kivet (1993), and Li and Maddala (1996), note that the menu of bootstrap-based test statistics extends beyond the simple pivotal/non-pivotal dichotomy. In this context we believe that is instructive to examine these issues in the design of bootstrap-based hypothesis tests.

2.1 Bootstrap Design

For the purposes of discussion we first introduce some additional notation. Let $\boldsymbol{\chi} = (x_1, ..., x_n)$ denote a random sample of size n drawn from $F(x;\boldsymbol{\theta})$ where F represents the population distribution function. For simplicity we assume that $F(x;\boldsymbol{\theta}) = N(\mu, \sigma^2)$ with $\boldsymbol{\theta} = \{\mu, \sigma^2\}$. Using this data we want to find an estimate of the sample mean, $\widehat{\mu}$, and evaluate its accuracy as an estimate of μ. If we wish to conduct hypothesis tests of the form $H_0 : \mu = \mu_0$ then a test statistic such as

$$Q = Q(\boldsymbol{\chi}) = n^{\frac{1}{2}}(\widehat{\mu} - \mu_0)/\sigma, \tag{1}$$

whose distribution under H_0 is (asymptotically) independent of $\boldsymbol{\theta}$, will prove useful. If we use bootstrap procedures then for $\boldsymbol{\chi}_r^* = (x_1^*, ..., x_n^*)'$ equal to the *rth* artificial sample of size n (conditional on the *fitted* null model) from $F(x;\hat{\boldsymbol{\theta}})$, the sample distribution function of the test statistic, then $Q^* = Q(\boldsymbol{\chi}^*)$ is the bootstrap analog of Q, with $Q_r^* = Q(\boldsymbol{\chi}_r^*)$ denoting the associated test statistic for the *rth* replication. The distribution function of Q^*, denoted $H(., \widehat{\boldsymbol{\theta}})$, is the bootstrap *estimator* for the null distribution[6] of Q, $H(., \boldsymbol{\theta})$. An estimate of $H(., \widehat{\boldsymbol{\theta}})$ is based upon R independent realisations $Q_1^*, ..., Q_{R.}^*$.

The importance of bootstrap design has been discussed by Hall and Wilson (1991). For example, in testing H_0 against $H_1 : \mu \neq \mu_0$, the authors note that the bootstrap distribution of $Q^* = \sqrt{n}(\hat{\mu}^* - \hat{\mu})/\hat{\sigma}^*$ is a better approximation to the distribution of $Q = \sqrt{n}(\hat{\mu} - \mu_0)/\hat{\sigma}$ under H_0 than $S^* = \hat{\mu}^* - \hat{\mu}$ is to the

[6] We suppress the dependence of the distributions on n in order to minimise notational burden.

distribution of $S = \hat{\mu} - \mu_0$. This follows since the asymptotic distribution of both S and S^* are scale dependent where the magnitude of the scale is likely to be different. The greater this difference the higher the error in rejection probability since S^* fails to approximate S. Similarly it is recommended that studentizing be achieved using $\sqrt{n}(\hat{\mu}^* - \hat{\mu})/\hat{\sigma}^*$ rather than $\sqrt{n}(\hat{\mu}^* - \hat{\mu})/\hat{\sigma}$. This is based upon a recognition that in the former test statistic $\widehat{\sigma}^*$ is a better estimate of the standard deviation of $\widehat{\mu}^*$. Additionally, as Hall (1992) notes, the design of the bootstrap affects the rejection probability of the test under H_1 as well as under H_0. When $\hat{\mu}$ is used to construct the bootstrap samples then the bootstrap test statistic should be $\sqrt{n}(\hat{\mu}^* - \hat{\mu})/\hat{\sigma}^*$ rather than $\sqrt{n}(\hat{\mu}^* - \mu_0)/\hat{\sigma}^*$ since the latter statistic leads to a procedure with zero local asymptotic power.

We also emphasise that pivoting can be accomplished *exactly* only in simple cases (i.e. as above). In all other cases, the pivoting is asymptotic and in this sense we refer to *asymptotically pivotal tests.*

2.2 Simulation version Bootstrap Error

Brown (2000), in noting the limitations of bootstrap procedures, draws an important distinction between simulation and bootstrap approximation error. The distinction between these two types of error has also been made by Davison and Hinkley (1997). The two types of error represent the total error in constructing the exact finite sample distribution of the statistic of interest.

Bootstrap or statistical error derives from resampling from the distribution function $F(x;\widehat{\boldsymbol{\theta}})$, rather than the population distribution function $F(x;\boldsymbol{\theta})$. For large n, $F(x;\widehat{\boldsymbol{\theta}}) \propto F(x;\boldsymbol{\theta})$, thereby guaranteeing that the distribution of $Q(\boldsymbol{\chi}^*)$ is close to that of $Q(\boldsymbol{\chi})$. Simulation error will be introduced since the estimate of the bootstrap distribution of $Q(\boldsymbol{\chi}^*)$ will be obtained from a finite number of replications. For example, with the true distribution of the test statistic denoted by $H(x;\boldsymbol{\theta})$, and the bootstrap distribution by $H(x;\widehat{\boldsymbol{\theta}})$, the difference between the two is the bootstrap error. An estimate of the bootstrap distribution of $H(x;\widehat{\boldsymbol{\theta}})$ is provided by $\frac{1}{R}\sum\limits_{r=1}^{R} \mathbf{1}(Q(\boldsymbol{\chi}_r^*) < x)$[7] and the difference between these two distributions is the simulation error which tends to zero as R tends to infinity. Brown (2000) notes that in many applications simulation

[7] $\mathbf{1}(.)$ denotes the indicator function.

approximation error can be significant and may dominate bootstrap error. Though the relative importance of simulation error is likely to vary across applications, a number of general comments can be made. First, it is likely that if the estimated model depends upon a large number of covariance parameters, then simulation error will be large. Second, the construction of a pivotal test statistic based upon a *double* bootstrap along the lines of Beran (1988) and Coulibaly and Brorsen (1997), will introduce compounded simulation error. Brown (2000) shows that for pivotal statistics generated by a nested bootstrap, it is necessary that $R \geqslant n^2$ (n^3) for the inner (outer) loop, giving a total number of simulations equal to n^5.

3 The Structure of the Cox Test Statistic

The essence of the Cox non-nested test is that the mean adjusted ratio of the maximised log-likelihoods of two non-nested models has a well defined *limiting* distribution under the null hypothesis. Below we introduce notation and present the form of the Cox test statistic. We do not derive the limiting distribution of the test statistic since this has been done elsewhere (see Pesaran (1987b)).[8]

3.1 Preliminaries

First, we denote two rival (conditional) nonnested models by

$$\begin{aligned} H_f &: \mathcal{F}_\theta = \{f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}, \quad i = 1, \ldots, T \\ H_g &: \mathcal{F}_\lambda = \{g(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\lambda}), \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}, \end{aligned}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are respectively k_f and k_g vectors of unknown parameters belonging to the non-empty compact sets $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}$, and where $\mathbf{x}$ and $\mathbf{z}$ represent the conditioning variables. For the sake of notational simplicity we shall also use $f_i(\boldsymbol{\theta})$ and $g_i(\boldsymbol{\lambda})$ in place of $f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})$ and $g(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\lambda})$, respectively.

Both $f(.)$ and $g(.)$ define a family of distributions over the respective parameter spaces, $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}$. Specifying H_f to be the null hypothesis, $f(y|\boldsymbol{\theta}_0)$

[8] For further discussion on non-nested tests and in particular the Cox test statistic, see Pesaran and Weeks (2000).

and $g(y|\boldsymbol{\lambda}(\boldsymbol{\theta}_0))$ denote the respective true and pseudo-true models. Assuming an independent and identically distributed sample of n observations, the log-likelihood for the respective samples may be written $l_f(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f_i(\boldsymbol{\theta})$ and $l_g(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} \log g_i(\boldsymbol{\lambda})$.

If $f(.)$ and $g(.)$ are non-nested densities then the expectation

$$E_f[l_f(\boldsymbol{\theta}) - l_g(\boldsymbol{\lambda})], \tag{2}$$

does not evaluate to zero. Cox (1961,1962) proposed a procedure such that a modified log-likelihood ratio has a well-defined limiting distribution. The use of this statistic in applied work has been limited to a restricted number of applications due to two principal problems. First, in order to estimate (2) we require a consistent estimate of the pseudo true value, $\boldsymbol{\lambda}(\boldsymbol{\theta}_0)$. Second, in most cases even given such an estimate, the expectation will still be intractable[9].

Using the notation set out above we may write the numerator of the Cox test statistic as

$$T_f = l_f(\widehat{\boldsymbol{\theta}}) - l_g(\widehat{\boldsymbol{\lambda}}) - C_{fg}(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\lambda}}), \tag{3}$$

The last term on the right-hand side of (3), $C_{fg}(\widehat{\theta}, \widetilde{\boldsymbol{\lambda}})$, represents a consistent estimator of $C_{fg}(\boldsymbol{\theta}_0, \boldsymbol{\lambda}(\boldsymbol{\theta}_0))$, the KL measure of closeness of $f(.)$ and $g(.)$ under f which is equal to the expectation in (2). This may be written as $C_{fg}(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\lambda}}) = E_f[l_f(\widehat{\boldsymbol{\theta}}) - l_g(\widetilde{\boldsymbol{\lambda}})]$, and is an estimator of the difference between the expected value of the two maximised log-likelihoods under the distribution given by $f(.)$; $\widetilde{\boldsymbol{\lambda}}$ is any consistent estimator for $\lambda(\boldsymbol{\theta}_0)$. Weeks (1998) in testing probit and logit models of discrete choice, distinguished between three variants, $\widetilde{\boldsymbol{\lambda}} = \{\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^R(\widehat{\boldsymbol{\theta}}), \overline{\boldsymbol{\lambda}}\}$: $\widehat{\boldsymbol{\lambda}}$ is the observed pseudo maximum likelihood estimator (MLE), $\boldsymbol{\lambda}^R(\widehat{\boldsymbol{\theta}}) = 1/R\sum_{r=1}^{R} \widehat{\boldsymbol{\lambda}}^r(\widehat{\boldsymbol{\theta}})$ is a simulation-based estimator where $\widehat{\boldsymbol{\lambda}}^r(\widehat{\boldsymbol{\theta}})$ represents the solution to $Argmax_{\lambda}\{L_g^r(\boldsymbol{\lambda})\}$ where $L_g^r(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \ln g(\mathbf{y}_i^r(\hat{\boldsymbol{\theta}})|\mathbf{z}_i, \boldsymbol{\lambda})$, $\mathbf{y}_i^r(\hat{\boldsymbol{\theta}})$ is the rth draw of $\mathbf{y}_i$ under H_f using $\widehat{\boldsymbol{\theta}}$, and R is the number of simulations. Note that for both $n \to \infty$ and $T \to \infty$ then $\operatorname{plim} \boldsymbol{\lambda}_R(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\lambda}(\boldsymbol{\theta}_0)$. $\overline{\boldsymbol{\lambda}}$ is due to Kent (1986) and is an estimator derived from maximising the fitted log-likelihood.

[9] An exception is the application of the Cox test to both binary and multinomial probit and logit models. Independent of the dimension of the choice set, the expected difference between the two log-likelihoods under the null has a relatively simple, closed form expression.

In this study we utilise bootstrap procedures to construct the *empirical distribution function* of the Cox test statistic, and simulation methods to evaluate the expectation in (2) applying these methods to testing linear versus loglinear models, two versus three regime SETAR models, and SETAR trend versus EDTAR trend models. For all cases, the Kullback-Leibler measure of closeness cannot be derived analytically.

A simulation-based estimator of $C_{fg}(\widehat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}})$ has been suggested by Pesaran and Pesaran (1993) and is given by

$$C_{fg,R}(\widehat{\boldsymbol{\theta}}, \lambda^R(\widehat{\boldsymbol{\theta}})) = \frac{1}{R}\sum_{r=1}^{R}(l_f(\widehat{\boldsymbol{\theta}}) - l_g(\boldsymbol{\lambda}^R(\widehat{\boldsymbol{\theta}})). \tag{4}$$

Obviously the use of bootstrap testing procedures using a non-pivotal statistic does not require the mean adjustment facilitated by (4). However, *pivotal* (or bootstrap-t) procedures require both mean and variance adjustments in order to effect asymptotic pivotalness. In this context (4) represents one approach to centring the log-likelihood ratio statistic, whereby both $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\lambda}^R(\widehat{\boldsymbol{\theta}})$ are treated as fixed parameters. An alternative method of mean adjustment is given by the following estimator of KLIC

$$C_{fg,R}(\widehat{\boldsymbol{\theta}}^1, \ldots, \widehat{\boldsymbol{\theta}}^R, \widehat{\boldsymbol{\lambda}}^1(\widehat{\boldsymbol{\theta}}), \ldots, \widehat{\lambda}^R(\widehat{\boldsymbol{\theta}})) = \frac{1}{R}\sum_{r=1}^{R}(l_f(\widehat{\boldsymbol{\theta}}^r) - l_g(\widehat{\boldsymbol{\lambda}}^r(\widehat{\boldsymbol{\theta}})), \tag{5}$$

where the parameter arguments to both $l_f(.)$ and $l_g(.)$ are allowed to vary across each rth replication.[10] We examine the small sample properties of the Cox test statistic using both (4) and (5).

In examining the variance of the limiting distribution of $\sqrt{n}T_f$ under H_f, denoted $v_f(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0)$, we utilise a decomposition employed by Orme (1994)

$$v_f = var(M) - cov(D, U)(var(U))^{-1}cov(D, U)', \tag{6}$$

where $var(M)$ represents the variance of the observed log-likelihood ratio and D and U are given by

$$D \;=\; l_f(\boldsymbol{\theta}_0) - l_g(\boldsymbol{\lambda}(\boldsymbol{\theta}_0))$$

[10] See Coulibaly and Brorsen (1997).

$$U = \frac{\partial}{\partial \boldsymbol{\theta}} l_f(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}_0}. \tag{7}$$

As Pesaran and Pesaran (1995) note, the second term on the right hand side of (6) represents the sampling uncertainty associated with the parameters estimated under the null. In an application of the Cox test procedure to a test linear and log-linear models, the authors consider three AE versions of v_f. Two of these exploit the information equality: an outer-product estimator calculating the variance of U using

$$E_f[\{\frac{\partial}{\partial \boldsymbol{\theta}} \log f(y|\boldsymbol{\theta}_0)\frac{\partial}{\partial \boldsymbol{\theta}} \log f(y|\boldsymbol{\theta}_0)'\}]; \tag{8}$$

and an estimator using

$$-E_f\{\frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\}. \tag{9}$$

The advantage of (8) is that it only requires evaluation of the vector of scores (for each sample point). This is particularly important in the case of highly non-linear models where the evaluation of the matrix of second derivatives of the log-likelihood is especially burdensome.[11] The disadvantage of this method is the well known poor finite sample properties of variance estimators based upon the outer-product of the gradient (see Davidson and MacKinnon (1981)). We also note that the estimate of the variance based on (9) may be negative in small samples.

A third AE of the variance of the Cox test statistic utilises only the first term in (6), and thus ignores the variance component due to the sampling uncertainty of the estimated parameters under the null model. In a comparison of the performance of the Cox test using these three different estimators for the variance, Pesaran and Pesaran find that this particular version exhibits superior performance relative to estimators based upon an OPG and an observed Hessian estimator.

We also consider a simulation-based estimator of the variance which utilises the information contained in the R replications used to mean adjust the Cox test statistic as in (4) and (5). Note that by using the R components $l_f(\widehat{\boldsymbol{\theta}}^r) -$

[11] In addition, many optimisation routines that are commonly used in these models (i.e. E.R. Berndt (1974)) rely solely upon the gradient of the log-likelihood.

$l_g(\widehat{\boldsymbol{\lambda}}^r(\widehat{\boldsymbol{\theta}}))$ we can construct a variance estimator using

$$\widehat{V}^2_{sim} = \sum_{i=1}^{R} \left\{ (l_f(\widehat{\boldsymbol{\theta}}^r - l_g(\widehat{\boldsymbol{\lambda}}^r(\widehat{\boldsymbol{\theta}}))) - \frac{1}{R} \sum_{r=1}^{R} l_f^r(\widehat{\boldsymbol{\lambda}}^r(\widehat{\boldsymbol{\theta}})) - l_g^r(\widehat{\boldsymbol{\lambda}}^r(\widehat{\boldsymbol{\theta}}))) \right\}^2 /R - 1. \tag{10}$$

(10) represents a numerical estimator of the variance using the set of simulated log-likelihood ratios which were used to construct the numerator of the Cox test statistic using either (4) or (5).

3.2 Resampling the Likelihood Ratio Statistic

Utilising a parametric bootstrap we present below a simple algorithm for resampling the likelihood ratio statistic which we use to construct the empirical distribution function of the test statistic. For the purpose of exposition the algorithm is presented for the non-pivotal bootstrap.

1. Generate R samples of size n by sampling from the *fitted* null model $f(\widehat{\boldsymbol{\theta}})$. y_i^r denotes the ith observation for the rth bootstrap-sample.

2. For each rth sample the pair $(\widehat{\boldsymbol{\theta}}^r, \widehat{\boldsymbol{\lambda}}^r)$ represent the parameter estimates obtained by maximising the log likelihoods.

   $$l_f^r(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i^r(\widehat{\boldsymbol{\theta}})|\boldsymbol{\theta}), l_g(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} \log g(y_i^r(\widehat{\boldsymbol{\theta}})|\boldsymbol{\lambda}). \tag{11}$$

   The simulated log likelihood ratio statistic is then

   $$T_f^r = l_f^r(\widehat{\boldsymbol{\theta}}^r) - l_g^r(\widehat{\boldsymbol{\lambda}}^r).$$

3. By constructing the empirical cdf we can compare the *observed* test statistic, $T_f = l_f(\widehat{\boldsymbol{\theta}}) - l_g(\widehat{\boldsymbol{\lambda}})$, with critical values obtained from the R independent (conditional) realisations of T_f^r. The p-value obtained from the bootstrap procedure is then given by[12]

   $$p = \frac{1 + \sum\limits_{r=1}^{R} \mathbf{1}(T_f^r \geq T_f)}{R+1}. \tag{12}$$

[12] If T is discrete then repeat values of T can occur requiring that we make an adjustment to (12).

The bootstrap procedure outlined above simply resamples the likelihood ratio statistic *without* pivoting. As discussed in Section 2 there are a number of alternative test statistics which are conjectured to represent an improvement over classical first order methods. Table 1 summarises the set of test statistics used in both empirical applications. We note that Table 1 includes two pivotal bootstrap test statistics P and D. P (D) is based upon a single (double) bootstrap design. To highlight the distinction between these two designs we consider the following algorithm.

Outer loop

Let $\mathbf{y}^1, ..., \mathbf{y}^R$ denote R bootstrap samples of size n, each conditional upon the *fitted* null model $f(\widehat{\boldsymbol{\theta}})$. P^r denotes an estimate of the modified likelihood ratio statistic using the rth bootstrap sample. Note that for each rth replication the mean adjustment in P is the same as P^r. Therefore although the scalar mean adjustment $\tau = \frac{1}{R}\sum[l_f^r - l_g^r]$ represents an unbiased estimate of $E_f[l_f - l_g]$ conditional upon $\widehat{\theta}$, it is not unbiased for $E_f^r[l_f^r - l_g^r]$. This is because the *observed* component of the bootstrap test statistic P^r is constructed using parameters $\widehat{\boldsymbol{\theta}}^r$ and $\widehat{\boldsymbol{\lambda}}^r$ such that a consistent estimate of the KL measure is $\alpha^r = \frac{1}{L}\Sigma[l_f^l - l_g^l]$ where l indexes a second round of L replications. Unlike τ, α^r varies across replications.

Inner loop

For each rth bootstrap sample, let $\mathbf{y}^{r_1}, ...\mathbf{y}^{r_L}$ represent L additional bootstrap samples, conditional upon the simulated *fitted* null model (i.e. using $f(\widehat{\boldsymbol{\theta}}^r)$). l_f^{rl} (l_g^{rl}) in D^r are, respectively, the mean log-likelihood (pseudo log-likelihood) for the lth inner-loop conditional upon $\widehat{\boldsymbol{\theta}}^r$.

The principal distinction between the test statistics P and D is that for the latter the construction of the bootstrap statistic D^r replicates the construction of the test statistic exactly. In fact, the distinction between them is similar to the distinction between the T and T^* test statistics presented in Section 2.1. Given the theoretical arguments presented, D should be preferred over P. However, there is an obvious trade-off between any benefits accruing to the size and power properties and the computational cost of a double bootstrap.

Table 1: Likelihood Ratio Tests: Bootstrap and Asymptotic

Bootstrap Tests		
	Observed Test Statistic	Bootstrap Test Statistic
Non-Pivotal	$T = l_f - l_g$	$T^r = l_f^r - l_g^r$
Pivotal (single)	$P = \frac{\sqrt{n}\{l_f - l_g - \frac{1}{R}\sum[l_f^r - l_g^r]\}}{\sqrt{\widehat{V}_s}}$	$P^r = \frac{\sqrt{n}\{(l_f^r - l_g^r) - \frac{1}{R}\sum[l_f^r - l_g^r]\}}{\sqrt{\widehat{V}_s^r}}$
Pivotal (double)	$D = \frac{\sqrt{n}\{l_f - l_g - \frac{1}{R}\sum[l_f^r - l_g^r]\}}{\sqrt{\widehat{V}_s}}$	$D^r = \frac{\sqrt{n}\{(l_f^r - l_g^r) - \frac{1}{L}\sum[l_f^{rl} - l_g^{rl}]\}}{\sqrt{\widehat{V}_s^r}}$
Studentised	$S = \frac{\sqrt{n}\{l_f - l_g\}}{\sqrt{\widehat{V}_s}}$	$S^r = \frac{\sqrt{n}\{l_f^r - l_g^r\}}{\sqrt{\widehat{V}_s^r}}$
Asymptotic Tests		
$\tilde{A} = \frac{\sqrt{n}\{l_f - l_g - \frac{1}{R}\sum[\tilde{l}_f^r - \tilde{l}_g^r]\}}{\sqrt{\widehat{V}_j}}$	$A = \frac{\sqrt{n}\{l_f - l_g - \frac{1}{R}\sum[l_f^r - l_g^r]\}}{\sqrt{\widehat{V}_j}}$	$j = op, s, sim$
$l_f = \frac{1}{n}\sum \ln f_i(y_i \vert x_i, \widehat{\boldsymbol{\theta}})$; $l_g = \frac{1}{n}\sum \ln g_i(y_i \vert z_i, \widehat{\boldsymbol{\lambda}})$ $l_f^r = \frac{1}{n}\sum \ln f_i(y_i^r(\widehat{\boldsymbol{\theta}}) \vert x_i, \widehat{\boldsymbol{\theta}}^r))$; $l_g^r = \frac{1}{n}\sum \ln g_i(y_i^r(\widehat{\boldsymbol{\theta}}) \vert z_i, \widehat{\boldsymbol{\lambda}}^r(\widehat{\boldsymbol{\theta}})))$ $\tilde{l}_f^r = \frac{1}{n}\sum \ln f_i(y_i^r(\widehat{\boldsymbol{\theta}}) \vert x_i, \widehat{\boldsymbol{\theta}}))$; $\tilde{l}_g^r = \frac{1}{n}\sum \ln g_i(y_i^r(\widehat{\boldsymbol{\theta}}) \vert z_i, \boldsymbol{\lambda}^R(\widehat{\boldsymbol{\theta}})))$ $l_f^{rl} = \frac{1}{n}\sum \ln f_i(y_i^{rl}(\widehat{\boldsymbol{\theta}}^r) \vert x_i, \widehat{\boldsymbol{\theta}}^{rl}))$; $l_g^{rl} = \frac{1}{n}\sum \ln g_i(y_i^{rl}(\widehat{\boldsymbol{\theta}}^r) \vert z_i, \widehat{\boldsymbol{\lambda}}^{rl}(\widehat{\boldsymbol{\theta}}^r)))$		

Following the discussion in Section 3.1 we also consider a number of alternate estimators for the variance of the Cox test statistic for the asymptotic tests. In the tables that follow $\widehat{V}_{op}$ denotes an OPG estimator for the variance, $\widehat{V}_s$ denotes a naive estimator which ignores parameter uncertainty under the null and $\widehat{V}_{sim}$ is the simulation estimator given in (10). Further, the asymptotic tests A and $\widetilde{A}$ refer, respectively, to equations (4) and (5). In test A the parameter arguments to the estimate of KLIC are constant over the R simulations where in $\widetilde{A}$ the parameters are specific to each rth replication.

4 A Test of the Linear versus Log-Linear Regression Model

In this section we evaluate the relative performance of a number of bootstrap tests, specifically T, P, and S, against two asymptotic tests: A_s and A_{sim} The experimental design used has been used previously by Godfrey, McAleer, and McKenzie (1988) and Pesaran and Pesaran (1995).[13] Using the notation defined above, we let H_f (H_g) denote the linear (logarithmic) regression model. When the linear model is true observations on the n pairs $\{y_t, x_t\}$ are generated according to

$$y_t = 500 + 5x_t + u_{tf},$$

where $u_{tf} \sim N(0, \sigma^2)$ and the autoregressive process x_t is defined by

$$x_t - 100 = 0.9(x_{t-1} - 100) + e_{tf},$$

with $e_{tf} \sim N(0, \sigma_f^2)$. Under the log-linear specification we have

$$\log y_t = 4.6 + 0.5 \log x_t + u_{tg},$$

where $u_{tg} \sim N(0, \eta^2)$. The autoregressive process x_t is defined by

$$\log(x_t/100) = 0.9 \log(x_{t-1}/100) + e_{tg},$$

[13] The experiments used in these two studies are extensions of those used by Aneuryn Evans and Deaton (1980).

with $e_{tg} \sim N(0, \eta_g^2)$. We vary sample size (n) using $n = (20, 40, 100)$, with the number of dgp replications set at 500, and the number of bootstrap replications (B) fixed at 200.

Across both sets of experiments, with either the linear or log-linear model serving as the null model, a number of general observations can be made. First, in all cases we observe the expected convergence of power for sample size of 100. Second, across both asymptotic and bootstrap tests, deviations of empirical from nominal size was less and power greater for experiments with lower values of σ_f^2 and σ_g^2. In general we observe reasonable performance for all bootstrap tests T (non-pivotal), P (pivotal) and S (studentised). Out of a total of 48 experiments (across different null models, sample sizes and $\sigma_{f(g)}^2$ = 8, 16), only in three instances do we observe a significant departure from nominal size. In terms of the asymptotic tests, the performance of the test statistic A_{sim}, with the variance calculated using a simulation estimator based upon (10)), is particularly noteworthy. With only one significant difference between nominal and empirical size, relative to all other test statistics which were correctly sized, A_{sim} consistently demonstrates higher power for small ($n < 80$) sample size.[14]

5 Non-Linear Threshold Models

In this section we present and briefly discuss alternative threshold models which will be used in the Monte Carlo experiments investigating the performance of asymptotic and bootstrap tests of nonnested hypotheses in a non-linear framework. Two testing situations will be considered. The first will concentrate on self-exciting threshold autoregressive (SETAR) and the second on endogenous delay threshold autoregressive (EDTAR) models.

5.1 Self-exciting threshold autoregressive models

The canonical form of a SETAR model with m regimes belonging to the class of threshold autoregressive (TAR) models, introduced and analysed extensively

[14] For example, compare 0.328 with comparable boostrap power values of $T = 0.264$, $S = 0.234$, and $P = 0.226$.

by Tong (1978, 1983, 1995), for a stochastic process $\{x_t\}$ is

$$x_t = \phi_{J_t,0} + \Phi_{J_t}(L)x_{t-1} + \sigma_{J_t}\epsilon_t, \quad t = p, \dots, T, \tag{13}$$

where $\Phi_{J_t}(L) = \sum_{k=1}^{p} \phi_{J_t,k} L^{k-1}$, $J_t = a'\mathbf{I}_t$, $a = (1, 2, \dots .m)'$, $\mathbf{I}_t = (\mathbf{1}(x_{t-d} \in A_1), \mathbf{1}(x_{t-d} \in A_2), \dots, \mathbf{1}(x_{t-d} \in A_m))'$, $A_i = [r_{i-1}, r_i)$, $i = 1, \dots, m$ where $\{r_1, \dots, r_{m-1}\}$ is a strictly increasing sequence of threshold parameters; $r_0 = -\infty$ and $r_m = \infty$ and $\mathbf{1}(.)$ denotes the indicator function. The sets, A_i, $i = 1, \dots, m$, define a partition of the real line, and d is referred to as the delay parameter. The basic idea is that the state of the system, at a specific point in the past, influences the current state of the system by regulating a switch between different linear laws of motion governing the system; these are referred to as regimes. SETAR models have been used to model macroeconomic series related to the business cycle (see, for example, Potter (1995)). Note that the representation in (13) treats all regimes of the SETAR model similarly. In some cases it is more intuitive to specify a *base* regime, say regime s, $0 < s \leq m$, and express the linear laws of motion of the other regimes as deviations from the base regime. Such a representation may be given by

$$x_t = \phi_{s,0} + \Phi_s(L)x_{t-1} + \Sigma_{i=1, i \neq s}^{m} \mathbf{1}(r_{i-1} \leq x_{t-d} < r_i)(\psi_{i,0} + \psi_i(L)x_{t-1}) + c_t\epsilon_t, \tag{14}$$

where $\psi_i = \phi_{i,0} - \phi_{s,0}$, $\Psi_i(L) = \Phi_i(L) - \Phi_s(L)$, $i = 1, \dots, m$, $i \neq s$ and $c_t = \Sigma_{i=1}^{m} \mathbf{1}(r_{i-1} \leq x_{t-d} < r_i)\sigma_i$. See also (18) below.

The first set of Monte Carlo experiments will compare a 2-regime and a 3-regime SETAR model[15]. Nonnested hypothesis testing is not the only alternative for evaluating the two models, as noted earlier. The intuitive definition

[15] The main problem which is specific to the application of nonnested hypothesis testing to threshold models is the discontinuity and/or non-differentiability associated with threshold parameters. The problem is not so serious since the higher rate of convergence of the estimated threshold parameters to their true values, implies that the analysis may assume that these parameters are known (see Chan (1993)). Note that the higher rate of convergence for the threshold parameter occurs in models with discontinuous conditional means. When the conditional mean is continuous but not differentiable everywhere, the threshold parameter is $\sqrt{T}$-consistent, at least for simple continuous threshold autoregressive models. However, in this case the parameter estimates including the threshold parameters are asymptotically normal (See Chan and Tsay (1998)). Then, it can be conjectured that the asymptotic results which form the basis of the Cox test procedure hold. However, the fact remains that, currently, rigorous proofs are available only under the assumption of known threshold parameters.

of nonnested hypothesis indicates that if $r_2 \to r_1$ then a 3-regime SETAR model nests a 2-regime SETAR model. In our setup this is not allowed since $r_2 > r_1$. Additionally, the two models are nested if the autoregressive coefficients of two regimes of the 3-regime SETAR model are equal. However, this would imply that one of the threshold parameters is not identified under a 3-regime model. Since the densities of the two models in a nonnested testing setting must be well defined (see, for example, Pesaran (1987a) or White (1982)) we choose to impose the restriction of different autoregressive coefficients for different regimes. Essentially, this is similar to the Davies problem (see Davies (1977)) arising in a number of linearity tests.

5.2 Endogenous delay threshold autoregressive models

The second set of Monte Carlo experiments will concentrate on the models developed in Kapetanios (1998) following the work by Beaudry and Koop (1993b) and Pesaran and Potter (1997). These models are based on ideas from theoretical nonlinear economic models of the business cycle and especially the floor and ceiling model of output by Hicks (1949, 1950). The basic premise of the Hicks model is that when output deviates from its long run trend value, dampening nonlinear forces come into effect and push output towards its trend value. The model was originally proposed to model output, but any series driven by the business cycle such as industrial production or imports may be modelled similarly. We will denote such a series by $\{y_t\}$ with the trend value[16] denoted by $\{y_t^\tau\}$. The models distinguish between three regimes. One regime holds when the series evolves near its trend value, this is referred to as the corridor regime. The other two regimes are activated when the series deviates from its trend value either downwards or upwards and are referred to as the floor and ceiling regimes. The following indicator functions are used to define the regimes

$$\text{Floor Regime} \quad I_{f\,t} = \mathbf{1}(y_t < y_t^\tau - r_f), \quad r_f > 0 \tag{15}$$

$$\text{Corridor Regime} \quad I_{\text{cor}\,t} = \mathbf{1}(I_{f\,t} + I_{c\,t} = 0), \tag{16}$$

[16] For a discussion on the specification of the trend series see Chapter 1 in Kapetanios (1998). For the purposes of this paper the trend will be estimated using a recursive Hodrick-Prescott filter.

$$\text{Ceiling Regime} \quad I_{c\,t} = \mathbf{1}(y_t > y_t^{\tau} + r_c), \quad r_c > 0. \tag{17}$$

The parameters r_f and r_c are threshold parameters. We can model a stationary transformation of the series, a natural choice in this framework being $y_t - y_t^{\tau}$ although Δy_t may be used to provide a link with the existing literature on threshold models. Both transformations are considered in Kapetanios (1998). For the Monte Carlo experiments in this paper Δy_t is used.

The first model we consider is a variant of a 3-regime SETAR model[17] given by

$$\Delta y_t = \phi_{\text{cor},0} + \Phi_{\text{cor}}(L)\Delta y_t + I_{f\,t-1}(\phi_{f,0} + \Phi_f(L)\Delta y_t) + I_{c\,t-1}(\phi_{c,0} + \Phi_c(L)\Delta y_t) + h_t\epsilon_t, \tag{18}$$

where $h_t = \sigma_{\text{cor}} I_{\text{cor}\,t-1} + \sigma_f I_{f\,t-1} + \sigma_c I_{c\,t-1}$. σ_{cor}, σ_f, σ_c are parameters to be estimated. $\{\epsilon_t\}$ is assumed to be an independent and identically distributed (i.i.d.) sequence of disturbances with zero mean and unit variance; $\Phi_{\text{cor}}(L)$, $\Phi_f(L)$ and $\Phi_c(L)$ are polynomials in the lag operator L with orders p_{cor}, p_f and p_c respectively. Note that the above specification allows for regime specific heteroscedasticity to account for possible changes in the variance of the series in different regimes[18]. However, the specifications for the floor and ceiling regimes in this model do not provide a natural representation for the dampening effects in the Hicks model given that such effects will only be present if the coefficients in $\Phi_f(L)$ and $\Phi_c(L)$ have the appropriate signs. A more suitable specification is provided by the class of EDTAR models which is discussed in Kapetanios (1998), Pesaran and Potter (1997) and Altissimo and Violante (1996). This class makes use of the following feedback variables to model the dampening effects

$$F_t = \sum_{i=0}^{p_r} \left[(y_{t-i}^{\tau} - r_f - y_{t-i}) \prod_{j=0}^{i} I_{f\,t-j} \right] \tag{19}$$

$$C_t = \sum_{i=0}^{p_e} \left[(y_{t-i} - y_{t-i}^{\tau} - r_c) \prod_{j=0}^{i} I_{c\,t-j} \right]. \tag{20}$$

[17] Note that if we specify $y_t^{\tau} = y_{t-1}$, then the model in (18) is a standard 3-regime SETAR model whose canonical form is given in (13).

[18] This pattern of conditional heteroscedasticity is referred to as Qualitative Threshold Autoregressive Conditional Heteroscedasticity (QTARCH) (see also, Altissimo and Violante (1996)).

Then, Δy_t is given by

$$\Delta_{y_t} = \phi_0 + \Phi(L)\Delta_{y_t} + \theta_f F_{t-1} + \theta_c C_{t-1} + h_t \epsilon_t. \tag{21}$$

Both the feedback variables are constructed to be either positive or zero. Each extra time period spent in the 'floor' or 'ceiling' regime leads to a rise in the value of F_t and C_t respectively. Therefore, the role of the feedback variables is to measure the dampening effects on the economy during contractions and expansions.

5.3 Test Results

The parameter values for the DGPs used in the construction of the Monte Carlo samples are presented in Tables 4 and 5. In both cases the autoregressive coefficients are chosen to lie in the stable region and take small absolute values so as to reduce estimate biases in small samples. In table 4 the parameter values are self-explanatory with the parameterisation of the variance consistent with homoscedasticity across regimes. In table 5, p_f, p_c and p_{cor}, together with p_r, p_e and p denote, respectively, the order of the polynomials for both the linear and non-linear components of the models. Note also that the parameterisation of the SETAR trend and EDTAR trend model allows for both a common linear and variance component, with respective parameters ϕ_i $i = 1, \ldots, m$ and $\phi_{cor,i}$ ϕ_i $i = 0, ..., m$. Results are presented for a nominal size of 5% (0.05). The results of the experiments are presented in Tables 6 to 9.

The results are encouraging for the simple bootstrap procedures. In the first set of experiments under the null of a 2 regime SETAR model, all but two experiments indicate that the actual size is not significantly different from the nominal size, as can be seen from the results in Table 6. The D procedure performs satisfactorily as well. Although A_{sim} has lower power than size at a sample size of 100 its performance improves for a sample size of 150. The remaining asymptotic tests A_{op}, A_s, $\widetilde{A}_{op}$, $\widetilde{A}_s$ and $\widetilde{A}_{sim}$ perform badly by either under-rejecting or over-rejecting the null. It is interesting to note that the non-pivotal bootstrap test T procedure has very low power in smaller samples.

When the null is a 3-regime SETAR model, the results in Table 7 indicate that the performance of the tests deteriorates. All asymptotic tests have very large actual sizes. For two asymptotic tests ($\widetilde{A}_{op}$ and $\widetilde{A}_s$), actual size exceeds power. In addition we note that the test statistic does not seem to exhibit the assumed convergence properties, at least for sample sizes between 50 and 200. We can provide some intuition for this result based upon the following argument. Given that the true model is a 3-regime SETAR and H_f (H_g) are 3 (2) regimes SETAR models, the variance of the log-likelihood ratio will be inflated. This follows from the results of Bai and Perron (1998) on estimating and testing linear models with structural breaks. Given that it is possible to think of thresholds and breaks as similar non-linearities depending upon how observations are ordered, we know that if we underestimate the number of true thresholds, we will obtain a consistent estimate, say $\widehat{r}^*$, of one of the two thresholds.[19] However, across samples the limit of $\widehat{r}^*$ will oscillate between the two true population threshold values, and as a result so will the variance of the mean parameters and the variance of log L_g. Further, this problem will persist in large samples, given that the threshold estimate will not converge to a single value. To examine this phenomenon a little closer we increased the sample size substantially, examining size for A_{op} and A_s with the sample set at 1500. The respective sizes of 0.977 and 0.969 indicates a worsening of the situation. Following this, and based on the conjecture that the estimation of nuisance (threshold) parameters may be a partial explanation, we repeated the same set of experiments, but fixing the threshold parameters at the known truth. Significance levels of $A_{op} = 0.824$ and $A_s = 0.777$ revealed that the estimation precision of threshold parameters was not the problem. Finally, we examined the size properties of a variant of A_{op} and A_s, by replacing $\widehat{\lambda}$ in the analytical estimator of the variance by $\lambda^R(\widehat{\theta})$, denoting the new test statistic $A_{op,s}$ and $A_{s,s}$. The results $(A_{op,s} = 0.102, A_{s,s} = 0.007)$ indicate that these test statistics appear to converge but there is still a significant difference between nominal and empirical size.[20]

The simple bootstrap tests (namely $T, S,$ and P) perform well having actual sizes less than 10 % for all experiments. Experiments using the D, A_{sim} and $\widetilde{A}_{sim}$ procedures indicate that these tests perform very badly and have extremely large actual sizes. Additionally, the simulated samples used for the

[19] This will depend upon which threshold parameter dominates in terms of maximising the log-likelihood.

[20] Note that there was no significant difference between nominal and empirical size for the test $A_{op,s}$ with the threshold parameters fixed at truth.

construction of the bootstrap statistics and the variance estimator $\widehat{V}_{sim}$, behave erratically and cannot be estimated as they have too few observations in one of the three regimes. It seems likely that the simulation error introduced by two sets of simulations is large. For these reasons we do not report results for these tests and conclude that they may not be suitable for the class of nonlinear models we consider.

In the second set of experiments presented in Tables 8 and 9, bootstrap tests perform satisfactorily with actual sizes of around 10%. Although the asymptotic tests A_s and $\widetilde{A}_s$ perform relatively well under the SETAR trend this is not so under the EDTAR trend null. Power under the null of an EDTAR model is quite low, especially for the T procedure. Overall, simple bootstrap tests perform consistently better than asymptotic tests. P and S perform comparably, possibly because of the effects of the simulation error discussed earlier.

6 Conclusion

In this paper we have considered a variety of simulation-based testing procedures for nonnested models including bootstrap procedures. We have seen that the framework of nonnested hypotheses can be extended to include situations where the models under consideration may be thought of as nested but cannot be straightforwardly handled as such. The test of a SETAR model with 2 regimes against a SETAR model with 3 regimes is an example of such a phenomenon. We have carried out an extensive Monte Carlo investigation of the small sample properties of the bootstrap procedures and compared them to procedures which although using simulation methods to evaluate the test statistic still rely on asymptotic approximations. We conclude that simple bootstrap procedures can provide significant improvements compared to both asymptotic procedures and more complicated bootstrap procedures which aim to incorporate second order corrections. It is clear that the simulation error involved in those procedures is larger than the gain from any possible reduction in the statistical error. By conducting experiments over types of models with different degrees of nonlinearity we determined that it is important to distinguish between testing over linear and log-linear models where the gain of using the bootstrap is moderate and highly nonlinear models where the asymptotic approximations are in general very poor. In such cases bootstrap based procedures provide more significant improvements.

References

Altissimo, F., and G. Violante (1996): "Persistence and Nonlinearity in US. GNP and Unemployment: An Endogenous Delay Threshold Model.," University of Pennsylvania.

Aneuryn Evans, G., and A. Deaton (1980): "Testing Linear versus Logarithmic Regression Models," *Review of Economic Studies*, pp. 257–291.

Bai, J., and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econmetrica*, 66(1), 47–78.

Beaudry, P., and G. Koop (1993a): "Do Recessions Permanently Change Output?," *Journal of Monetary Economics*, 31, 149–164.

——— (1993b): "Do Recessions permanently change output ?," *Journal of Monetary Economics*, 31, 149–164.

Beran, R. (1988): "Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements," *Journal of the American Statistical Association*, 83(403).

Brown, W. B. (2000): "Simulation Variance Reduction for Bootstrapping," in *Simulation-Based Inference: Methods and Applications*, ed. by B. Mariano, T. Schuermann, and M. Weeks. CUP, Cambridge.

Coulibaly, N., and B. Brorsen (1997): "A Monte Carlo Sampling Approach to Testing Nonnested Hypotheses: Monte Carlo Results," Dept. of Agriculture Economics, Oklahoma State University, USA.

Cox, D. (1961): "Tests of Separate Families of Hypothesis," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.*

Davidson, R., and J. MacKinnon (1981): "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica*, 49, 781–793.

Davies, R. B. (1977): "Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 64(2), 247–254.

Davison, A., and D. Hinkley (1997): *Bootstrap Methods and their Applications.* Cambridge University Press, Cambridge, UK.

Efron, B. (1979): "Bootstrap Methods: Another Look at the Jacknife," *Annals of Statistics*, 7, 1–26.

E.R. Berndt, B.H. Hall, R. H. J. H. (1974): "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, 3, 653–666.

Godfrey, L., M. McAleer, and C. McKenzie (1988): "Variable Addition and Lagrange Multiplier Tests for Linear and Logarithmic Regression Models," *Review of Economics and Statistics*, 70, 492–503.

Hall, P. (1986): "On the Number of Bootstrap Simulations Required to Construct a Confidence Interval," *The Annals of Statistics*, 14(4).

——— (1988): "Theoretical Comparison of Bootstrap Confidence Intervals," *Annals of Statistics*, 16, 927–953.

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

Hall, P., and S. Wilson (1991): "Two Guidelines for Bootstrap Hypothesis Testing," *Biometrics*, 47, 757–762.

Hartigan, J. (1986): "Comment on the Paper by Efron and Tibshirani," *Statistical Science*, 1, 75–76.

Horowitz, J. (1995): "Bootstrap Methods in Econometrics: Theory and Numerical Performance," manuscript, Department of Economics, University of Iowa.

Jeong, J., and G. Maddala (1993): "A Perspective on Application of Bootstrap Methods in Econometrics," *Handbook of Statistics*, 11, 573–605.

KAPETANIOS, G. (1998): "Essays on the Econometric Analysis of Threshold Models," Ph.D. Thesis, University of Cambridge.

KENT, J. (1986): "The Underlying Structure of Nonnested Hypothesis Tests," *Biometrika*, 7, 333–43.

KILIAN, L. (1998): "Pitfalls in Constructing Bootstrap Confidence Intervals for Asymptotically Pivotal Statistics," Dept. of Economics, University of Michigan.

LI, H., AND G. MADDALA (1996): "Bootstrapping Time Series Models," *Econometric Reviews*, 15(2), 115–158.

ORME, C. (1994): "Non-Nested Tests for Discrete Choice Models," .

PESARAN, H., AND M. WEEKS (2000): "Non-Nested Hypothesis Tests," in *Theoretical Econometrics*, ed. by B. Baltagi. Basil Blackwell, Oxford.

PESARAN, M. (1987a): "Global and Partial Nonnested Hypotheses and Asymptotic Local Power.," *Econometric Theory*, 3, 69–97.

PESARAN, M. H. (1987b): "Global and Partial Nonnested Hypotheses and Asymptotic Local Power," *Econometric Theory*, 3, 69–97.

PESARAN, M. H., AND B. PESARAN (1993): "A Simulation Approach to the Problem of Computing Cox's Statistic for Testing Non-Nested Models," *Journal of Econometrics*, 57, 377–92.

——— (1995): "A Non-Nested Test of Level Differences Versus Log-Differenced Stationary Models," *Econometric Reviews*, 14(2), 213–27.

PESARAN, M. H., AND S. POTTER (1997): "A Floor and Ceiling Model of U.S. Output," *Journal of Economic Dynamics and Control*, 21(4–5), 661–696.

POTTER, S. (1995): "A Nonlinear Approach to US GNP.," *Journal of Applied Econometrics*, 10, 109–125.

SINGH, K. (1981): "On the Asymptotic Accuracy of Efron's Bootstrap," *Annuals of Statistics*, 9, 1187–1195.

TIBSHIRANI, R. (1988): "Variance Stabilization and the Bootstrap," *Biometrika*, 75, 433–444.

TONG, H. (1978): "On a Threshold Model," in *Pattern Recognition and Signal Processing*, ed. by C. Chen. Sijthoff and Noordhoff, Amsterdam.

——— (1983): *Threshold Models in Nonlinear Time Series Analysis*. Springer Verlag.

——— (1995): *Nonlinear Time Series: A Dynamical System Approach*. Oxford University Press.

VINOD, H. (1993): "Bootstrap Methods: Applications in Econometrics," *Handbook of Statistics*, 11, 629–661.

WEEKS, M. (1996): "Testing the Binomial and Multinomial Choice Models Using Cox's Non-Nested Test," *Journal of the American Statistical Association (Papers and Proceedings).*

WHITE, H. (1982): "Regularity Conditions for Cox's Test of Nonnested Hypotheses," *Journal of Econometrics*, 19, 301–18.

Table 2: Bootstrap-Based Likelihood Ratio Tests

Linear Model (H_f) is the Data Generating Process					
		$\sigma_f = 8$		$\sigma_f = 16$	
Sample Size[a]	Test	Linear	Log-Linear	Linear	Log-Linear
20	T	0.040	0.054	0.066	0.100
	S	0.058	0.060	0.062	0.098
	P	0.054	0.060	0.076*	0.094
	A_s	0.134*	0.154	0.144*	0.274
	A_{sim}	0.042	0.068	0.052	0.078
40	T	0.052	0.264	0.042	0.580
	S	0.038	0.234	0.042	0.584
	P	0.064	0.226	0.076*	0.508
	A_s	0.126*	0.294	0.058*	0.626
	A_{sim}	0.036	0.328	0.044	0.624
80	T	0.052	0.376	0.052	1.000
	S	0.060	0.360	0.062	1.000
	P	0.056	0.314	0.046	0.994
	A_s	0.080*	0.428	0.056	0.996
	A_{sim}	0.056	0.420	0.044	1.000
100	T	0.048	0.548	0.046	1.000
	S	0.070	0.544	0.038	1.000
	P	0.056	0.502	0.046	1.000
	A_s	0.080*	0.560	0.062	1.000
	A_{sim}	0.046	0.582	0.042	1.000

[a] Starred entries indicate that the estimated size is significantly different from 0.05 at the 5% significance level. The variance of the estimated size is obtained using the normal approximation to the binomial distribution and is given by $N^{-1}\widehat{a}(1-\widehat{a})$ where N is the number of Monte Carlo replications and $\widehat{a}$ is the estimated size.

Table 3: Bootstrap-Based Likelihood Ratio Tests

Log-Linear Model (H_g) is the Data Generating Process

		$\sigma_g = 8$		$\sigma_g = 16$	
Sample Size^b	Test	Linear	Log-Linear	Linear	Log-Linear
20	T	0.034	0.046	0.200	0.044
	S	0.064	0.056	0.196	0.058
	P	0.058	0.052	0.186	0.068
	A$_s$	0.154	0.112*	0.350	0.182*
	A$_{sim}$	0.038	0.044	0.260	0.036
40	T	0.246	0.040	0.304	0.046
	S	0.246	0.050	0.304	0.094*
	P	0.236	0.044	0.288	0.042
	A$_s$	0.400	0.068	0.426	0.086*
	A$_{sim}$	0.256	0.034*	0.340	0.036
80	T	0.404	0.058	0.872	0.058
	S	0.376	0.048	0.914	0.054
	P	0.326	0.062	0.840	0.062
	A$_s$	0.454	0.088*	0.900	0.082*
	A$_{sim}$	0.418	0.056	0.882	0.066
100	T	0.424	0.068	1.000	0.038
	S	0.438	0.054	0.988	0.042
	P	0.390	0.078	0.990	0.042
	A$_s$	0.508	0.076*	1.000	0.046
	A$_{sim}$	0.456	0.056	1.000	0.034*

[b] See footnote (a) Table 2.

Table 4: DGPs for first set of Monte Carlo experiments

	DGP 1[a]	DGP 2[b]
p	2	2
r_1	0	-0.3
r_2		0.3
d	1	1
$\phi_{1,0}$	0.1	0.2
$\phi_{1,1}$	0.2	0.1
$\phi_{1,2}$	-0.1	-0.1
$\phi_{2,0}$	0.2	0.1
$\phi_{2,1}$	0.1	0.2
$\phi_{2,2}$	-0.2	0.2
$\phi_{3,0}$		0.2
$\phi_{3,1}$		0.2
$\phi_{3,2}$		-0.2
σ_1^2	1	2.25
σ_2^2	1	2.25
σ_3^2		2.25

[a] 2-regime SETAR model
[b] 3-regime SETAR model

Table 5: DGPs for second set of Monte Carlo experiments

	DGP 3[a]	DGP 4[b]
r_f	1	1
r_c	1	1
p_f	2	
p_c	2	
p_{cor}	2	
p_r		1
p_e		1
p		2
$\phi_{\text{cor},0}$	0.1	
$\phi_{\text{cor},1}$	0.2	
$\phi_{\text{cor},2}$	-0.2	
$\phi_{f,0}$	0	
$\phi_{f,1}$	-0.1	
$\phi_{f,2}$	-0.15	
$\phi_{c,0}$	0.05	
$\phi_{c,1}$	-0.05	
$\phi_{c,2}$	-0.1	
ϕ_0		0.1
ϕ_1		0.2
ϕ_2		-0.2
θ_f		0.5
θ_c		-0.5
σ_f^2	1	1
σ_c^2	1	1
σ_{cor}^2	1	1

[a] SETAR trend model
[b] EDTAR model

Table 6: Test size and power under H_f: 2-regime SETAR model against H_g: 3-regime SETAR model. True DGP for size: 2 regime SETAR, True DGP for power: 3 regime SETAR

Testing Procedures		Size[a]				Power			
		Sample Size				Sample Size			
		50	100	150	200	50	100	150	200
Asymptotic Testing Procedures[b]	A_{op}	0.709**	0.031**	0.013**	0.010**	0.770	0.925	1.000	1.000
	A_s	0.639**	0.018**	0.005**	0.004**	0.739	0.918	1.000	1.000
	$\tilde{A}_{op}$	0.998**	0.806**	0.541**	0.381**	0.997	0.995	1.000	1.000
	$\tilde{A}_s$	0.979**	0.580**	0.372**	0.258**	0.997	0.992	1.000	1.000
	A_{sim}	0.040	0.045	0.048	0.042	0.017	0.274	0.966	1.000
Bootstrap Testing Procedures[c]	T	0.040	0.039	0.054	0.045	0.075	0.268	0.976	1.000
	S	0.058	0.041	0.062	0.038*	0.772	0.988	1.000	1.000
	P	0.070*	0.040	0.059	0.044	0.376	0.958	1.000	1.000
	D	N/A	0.045	0.030	N/A	N/A	0.860	0.995	N/A

[a] Starred entries indicate that the estimated size is significantly different from 0.05 at the 5% significance level. Double stars indicate difference at the 1% significance level. The variance of the estimated size is obtained using the normal approximation to the binomial distribution and is given by $N^{-1}\widehat{\alpha}(1-\widehat{\alpha})$ where N is the number of Monte Carlo replications and $\widehat{\alpha}$ is the estimated size.

[b] The test statistics for the asymptotic testing procedures are given in Table 3.1.

[c] The test statistics and bootstrap test statistics for the bootstrap testing procedures are given in Table 3.1.

Table 7: Test size and power under H_f: 3-regime SETAR model against H_g: 2-regime SETAR model. True DGP for size: 3 regime SETAR, True DGP for power: 2 regime SETAR

		Size				Power			
Testing		Sample Size				Sample Size			
Procedures[a]		50	100	150	200	50	100	150	200
Asymptotic	A_{op}	0.858**	0.797**	0.802**	0.835**	0.892	0.956	0.997	0.997
Testing	A_s	0.807**	0.711**	0.733**	0.772**	0.917	0.968	0.989	0.996
Procedures	$\tilde{A}_{op}$	0.909**	0.938**	0.962**	0.961**	0.735	0.804	0.783	0.836
	$\tilde{A}_s$	0.890**	0.919**	0.948**	0.951**	0.601	0.659	0.543	0.609
	A_{sim}	0.115**	0.683**	0.694**	0.735**	0.253	0.849	0.963	0.984
Bootstrap	T	0.095**	0.063	0.073**	0.077**	0.737	0.928	0.985	0.996
Testing	S	0.042	0.039	0.029**	0.024**	0.498	0.873	0.973	0.987
Procedures	P	0.288**	0.087**	0.087**	0.095**	0.750	0.903	0.986	0.994

[a] See notes in Table 6

Table 8: Test size and power under H_f: SETAR trend model against H_g: EDTAR trend model. True DGP for size: SETAR trend, True DGP for power: EDTAR trend

		Size		Power	
Testing		Sample Size		Sample Size	
Procedures[a]		100	150	100	150
Asymptotic	A_{op}	0.435**	0.505**	0.805	0.910
Testing	A_s	0.130**	0.080	0.500	0.600
Procedures	$\tilde{A}_{op}$	0.190**	0.140**	0.685	0.790
	$\tilde{A}_s$	0.065	0.020**	0.410	0.470
	A_{sim}	0.030	0.090*	0.270	0.470
Bootstrap	T	0.100*	0.105*	0.470	0.640
Testing	S	0.100*	0.115**	0.455	0.640
Procedures	P	0.110**	0.135**	0.470	0.670

[a] See notes in Table 6

Table 9: Test size and power under H_f: EDTAR trend model against H_g: SETAR trend model. True DGP for size: EDTAR trend, True DGP for power: SETAR trend

Testing Procedures[a]		Size		Power	
		Sample Size		Sample Size	
		100	150	100	150
Asymptotic Testing Procedures	A_{op}	0.075	0.070	0.180	0.275
	A_s	0.055	0.035	0.075	0.150
	$\tilde{A}_{op}$	0.505**	0.370**	0.690	0.730
	$\tilde{A}_s$	0.315**	0.230**	0.450	0.545
	A_{sim}	0.03	0.035	0.065	0.075
Bootstrap Testing Procedures	T	0.055	0.045	0.065	0.100
	S	0.160**	0.100*	0.335	0.610
	P	0.115**	0.095*	0.215	0.370

[a] See notes in Table 6