*Econometrica*, Vol. 64, No. 2 (March, 1996), 413–430

# INFERENCE WHEN A NUISANCE PARAMETER IS NOT IDENTIFIED UNDER THE NULL HYPOTHESIS

BY BRUCE E. HANSEN[1]

Many econometric testing problems involve nuisance parameters which are not identified under the null hypotheses. This paper studies the asymptotic distribution theory for such tests. The asymptotic distributions of standard test statistics are described as functionals of *chi-square processes*. In general, the distributions depend upon a large number of unknown parameters. We show that a transformation based upon a conditional probability measure yields an asymptotic distribution free of nuisance parameters, and we show that this transformation can be easily approximated via simulation. The theory is applied to threshold models, with special attention given to the so-called *self-exciting threshold autoregressive model*. Monte Carlo methods are used to assess the finite sample distributions. The tests are applied to U.S. GNP growth rates, and we find that Potter's (1995) threshold effect in this series can be possibly explained by sampling variation.

KEYWORDS: Asymptotic theory, nonlinear models, thresholds, identification, $p$-values, hypothesis testing.

## 1. INTRODUCTION

THIS PAPER STUDIES THE PROBLEM of inference in the presence of nuisance parameters which are not identified under the null hypothesis. The models considered take the form of additive nonlinearity, allowing for stochastic regressors and weak dependence. The asymptotic distributions of standard tests are nonstandard and nonsimilar, which means (among other things) that tabulation of critical values is impossible. This paper proposes a conditional transformation which is analogous to an asymptotic $p$-value, and has an asymptotic uniform distribution under the null hypothesis. The transformation is not directly calculable, but can be approximated using simple simulation techniques.

There are many econometric hypotheses of interest with unidentified nuisance parameters. Some examples include: (i) common (canceling) ARMA roots; (ii) no ARCH effect in a GARCH or ARCH-M model; (iii) constancy of a regression coefficient with the alternative that it follows an AR(1) process; (iv) a single regime or state against the alternative of multiple regimes. For brevity, this paper does not explicitly examine these cases, although the methods of this paper can be extended to handle these cases as well.[2]

[1]This research was funded by a research grant from the National Science Foundation and a research fellowship from the Alfred P. Sloan Foundation. I gratefully acknowledge helpful conversations with Donald Andrews, Adrian Pagan, Simon Potter, and Halbert White. Insightful comments on earlier drafts by the co-editor and two referees led to a substantial rewriting of the paper.

[2]An earlier version of this paper (Hansen (1991)) used a more general notational framework which included nonlinear maximum likelihood estimation.

This paper concentrates on regression models with additive nonlinearity, which take the form $y_t = x_t'\alpha + h(z_t, \gamma)'\theta + \varepsilon_t$. Here, the null hypothesis of interest is whether the nonlinear term $h(z_t, \gamma)$ enters the regression. When $\theta = 0$ the parameter $\gamma$ is not identified. Specific examples of additive nonlinearity include the following. (i) Box-Cox transformations: $h(z_t, \gamma) = (z_t^\gamma - 1)/\gamma$; (ii) structural change: $h(z_t, \gamma) = \{t/n \le \gamma\} z_t$ where $\{\cdot\}$ is the indicator function; (iii) threshold models: $h(z_t, \gamma) = \{z_t \le \gamma\} z_t$; (iv) Bieren's (1990) consistent tests of functional form: $h(z_t, \gamma) = \exp(\gamma' z_t)$; (v) White's (1989) neural network tests of functional form: $h(z_t, \gamma) = \psi(\gamma' z_t)$, where $\psi(\cdot)$ is the logistic function.

There have been several recent papers on the subject of unidentified nuisance parameters which examine related issues. Andrews (1993b) analyzed tests for structural change. Andrews and Ploberger (1994) explore optimal testing but do not discuss methods to obtain critical values in practice. Andrews (1993a) discusses a range of econometric examples which suffer from the problem of unidentified nuisance parameters. King and Shively (1991) discuss the merits and difficulties of reparameterization as a means to handle the problem. Stinchcombe and White (1993) examine White's neural network tests. All of these papers investigate different aspects of the problem and should be viewed as complementary to the results of this paper.

In the next section, we introduce additive nonlinear regression and discuss pointwise and global test statistics. Section 3 introduces the conditional $p$-value transformation. Section 4 examines threshold regression, including the self-exciting threshold autoregressive (SETAR) model of Tong (1983). A Monte Carlo study is reported in Section 5. We find that the LM-based tests have excellent size and good power. In Section 6, we apply these tests to Potter's (1995) SETAR model of U.S. GNP and find that the apparent "significant" threshold effect may be explained by sampling error. Section 7 contains a brief conclusion. Proofs are left to an Appendix. Concerning notation, let $|A| = (\mathrm{tr}(A'A))^{1/2}$ denote the Euclidean norm of a matrix $A$, let $\|A\|_r = (E|A|^r)^{1/r}$ denote the $L^r$-norm of a random matrix, and let $\Rightarrow$ denote weak convergence with respect to the uniform metric.

## 2. FRAMEWORK AND TEST STATISTICS

The data is $\{w_t = (y_t x_t') : t = 1, \ldots, n\}$, which is a draw from some underlying probability space, and satisfies the regression relationship $y_t = x_{1t}'\beta_1 + h(x_t, \gamma)'\beta_2 + \varepsilon_t$, where $x_{1t}$ is a $k_1 \times 1$ subvector of the $k$ vector $x_t$, and $\varepsilon_t$ is a real-valued martingale difference sequence with respect to some increasing set of sigma-fields $\mathcal{F}_t$ to which $\varepsilon_t$ and $x_{t+1}$ are adapted, and $E\varepsilon_t^2 = \sigma^2 < \infty$. $\gamma$ takes values in $\Gamma$, a bounded subset of $R^q$, and the function $h(\cdot,\cdot)$ maps $R^k \times \Gamma$ into $R^m$. We will typically write $h(x_t, \gamma)$ more simply as $h_t(\gamma)$. It will be convenient to write the model in the alternate form

$$y_t = x_t(\gamma)'\beta + \varepsilon_t \tag{1}$$

where $x_t(\gamma) = (x_{1t}' h_t(\gamma)')'$ and $\beta = (\beta_1' \beta_2')'$.

The question of interest is whether the nonlinear term $h_t(\gamma)$ enters the regression, that is, whether $\beta_2 = 0$. The distributional theory will be facilitated by a local-to-null reparameterization: $\beta_2 = c/\sqrt{n}$. The null hypothesis is $H_0$: $c = 0$ with alternative $H_1: c \neq 0$. The test is nonstandard since $\gamma$ does not enter the regression (and therefore is not identified) under $H_0$.

Under $H_0$ the model simplifies to $y_t = x_{1t}'\beta_1 + \varepsilon_t$. The OLS estimators are $\tilde{\beta}_1 = (\Sigma_{t=1}^n x_{1t}x_{1t}')^{-1}(\Sigma_{t=1}^n x_{1t}y_t)$ and $\tilde{\sigma}_n^2 = \Sigma_{t=1}^n \tilde{\varepsilon}_t^2/(n-k_1)$, where $\tilde{\varepsilon}_t = y_t - x_{1t}'\tilde{\beta}_1$.

Under $H_1$, if $\gamma$ were known, then $\beta$ could be estimated by ordinary least squares on (1), yielding estimates $\hat{\beta}(\gamma) = (\Sigma_{t=1}^n x_t(\gamma)x_t(\gamma)')^{-1}(\Sigma_{t=1}^n x_t(\gamma)y_t)$, residuals $\hat{\varepsilon}_t(\gamma) = y_t - x_t(\gamma)'\hat{\beta}(\gamma)$, and sample variance $\hat{\sigma}_n^2(\gamma) = \Sigma_{t=1}^n \hat{\varepsilon}_t(\gamma)^2/(n-(k_1+m))$. When $\gamma$ is unknown, then its least-squares estimate can be found by minimization of $\hat{\sigma}_n^2(\gamma)$ over $\gamma \in \Gamma$, yielding $\hat{\gamma} = \operatorname{argmin} \hat{\sigma}_n^2(\gamma)$ and $\hat{\beta} = \hat{\beta}(\hat{\gamma})$. These are also the MLE when $\varepsilon_t$ is idd Gaussian. It will be useful to define the regression scores $s_t(\gamma) = x_t(\gamma)\varepsilon_t$ and their estimates under $H_0$ and $H_1$, respectively, $\hat{s}_t(\gamma) = x_t(\gamma)\hat{\varepsilon}_t(\gamma)$ and $\tilde{s}_t(\gamma) = x_t(\gamma)\tilde{\varepsilon}_t$.

If $\gamma$ were known, then the testing problem would not be complicated. A heteroskedasticity-robust Wald test takes the form

$$T_n(\gamma) = n\hat{\beta}(\gamma)'R\left(R\hat{V}_n^*(\gamma)R\right)^{-1}R'\hat{\beta}(\gamma)$$

where $R$ is the selector matrix $R = (0 I_p)'$, $\hat{V}_n^*(\gamma) = M_n(\gamma,\gamma)^{-1}\hat{V}_n(\gamma)M_n(\gamma,\gamma)^{-1}$, $\hat{V}_n(\gamma) = \frac{1}{n}\Sigma_{t=1}^n \hat{s}_t(\gamma)\hat{s}_t(\gamma)'$, and $M_n(\gamma_1,\gamma_2) = \frac{1}{n}\Sigma_{t=1}^n x_t(\gamma_1)x_t(\gamma_2)'$. Alternatively, a Lagrange multiplier statistic is found by setting $\hat{V}_n(\gamma) = \frac{1}{n}\Sigma_{t=1}^n \tilde{s}_t(\gamma)\tilde{s}_t(\gamma)'$. Tests which assume homoskedastic errors can be formed similarly.

If $\gamma$ were known *a priori*, then under conventional regularity conditions $T_n(\gamma)$ would have a "point-optimal" interpretation and an approximate $\chi_m^2$ null distribution in large samples. Hence $H_0$ can be tested by selecting a "reasonable" value of $\gamma$ a priori. We might expect this test procedure to work well if $\gamma$ is *known* a priori, or if the selected test statistic is not sensitive to the choice of $\gamma$. For many of the examples of interest, however, neither condition holds. The researcher is left with an unpleasant dilemma. Either $\gamma$ is selected in a completely arbitrary way (and thereby sacrifices power) or $\gamma$ is selected in some data-dependent fashion, in which case the chi-square distributional approximation will be invalid. In addition, researchers who estimate unrestricted models will be generating estimates of $\gamma$. They will be interested in the question: "Does the unrestricted model fit statistically better than the restricted model?" which cannot be answered by a point-optimal test. To avoid these difficulties, we turn to tests which do not require prior knowledge of $\gamma$.

Davies (1977, 1987) suggested testing $H_0$ by $\sup T_n = \sup_{\gamma \in \Gamma} T_n(\gamma)$, which equals the LR statistic when $T_n(\gamma)$ is the pointwise likelihood ratio statistic. Andrews and Ploberger (1994) examine tests of $H_0$ and argue that superior local power can be constructed from the statistics $\operatorname{ave} T_n = \int_\Gamma T_n(\gamma)\,dW(\gamma)$ and $\exp T_n = \ln(\int_\Gamma \exp(\frac{1}{2}Z(\gamma))\,dW(\gamma))$. All three statistics can be written as functions $g(T_n)$, where $g(\cdot)$ maps functionals on $\Gamma$ to $R$, and we write $T_n = \{T_n(\gamma): \gamma \in \Gamma\}$ as a random function on $\Gamma$. Each function $g$ is continuous with respect to the

uniform metric, monotonic in the sense that if $Z_1(\gamma) \leq Z_2(\gamma)$ for all $\gamma$ then $g(Z_1) \leq g(Z_2)$, and has the property that if $Z(\gamma) \to \infty$ for $\gamma$ for some subset of $\Gamma$ with positive $W$-measure, then $g(Z) \to \infty$. We will denote the test statistic as $g_n = g(T_n)$ for any choice of $g$.

Define $M(\gamma_1, \gamma_2) = E(x_t(\gamma_1)x_t(\gamma_2)')$ and $\bar{x}_t = \sup_{\gamma \in \Gamma} |x_t(\gamma)|$.

ASSUMPTION 1: *$w_t$ is strictly stationary and absolutely regular with mixing coefficients $\eta(m) = O(m^{-A})$ for some $A > v/(v-1)$ and $r \geq v > 1$; $E|\bar{x}_t|^{4r} < \infty$; $E|\varepsilon_t|^{4r} < \infty$; and $\inf_{\gamma \in \Gamma} \det(M(\gamma, \gamma)) > 0$.*

The absolute regular mixing coefficient $\eta(\mathscr{A}, \mathscr{B})$ between $\sigma$-fields $\mathscr{A}$ and $\mathscr{B}$ is defined as

$$\eta(\mathscr{A}, \mathscr{B}) = \frac{1}{2} \sup \sum_{(i,j) \in (I,J)} |P(A_i \cap B_j) - P(A_i)P(B_j)|,$$

where $A_i \subset A$, $B_j \subset B$, and the supremum is taken over all the finite partitions $(A_i)_{i \in I}$ and $(B_j)_{j \in J}$ respectively $\mathscr{A}$ and $\mathscr{B}$ measurable. Absolute regularity was first defined by Volkonskii and Rozanov (1959), and is stronger than strong mixing yet weaker than uniform mixing. Pham and Tran (1985) have shown that a wide class of linear processes with iid innovations (such as ARMA processes) are absolutely regular when the innovation has a bounded, continuous density, and thus these processes satisfy Assumption 1. Define $K_n(\gamma_1, \gamma_2) = \frac{1}{n}\sum_{t=1}^n s_t(\gamma_1)s_t(\gamma_2)'$, and $K(\gamma_1, \gamma_2) = E(s_t(\gamma_1)s_t(\gamma_2)')$. Under the stated moment conditions, we see that for all $\gamma_1, \gamma_2 \in \Gamma$, $M_n(\gamma_1, \gamma_2) \to M(\gamma_1, \gamma_2)$ and $K_n(\gamma_1, \gamma_2) \to K(\gamma_1, \gamma_2)$ a.s. We will need the stronger requirement of uniform convergence. At this point we give high-level conditions, supplying primitive conditions in Section 4.

ASSUMPTION 2: *For some $B < \infty$ and $\lambda > 0$, $\|(h_t(\gamma) - h_t(\gamma'))\varepsilon_t\|_{2v} \leq B|\gamma - \gamma'|^{\lambda}$.*

ASSUMPTION 3: *$M_n(\gamma_1, \gamma_2)$ and $K_n(\gamma_1, \gamma_2)$ converge almost surely to $M(\gamma_1, \gamma_2)$ and $K(\gamma_1, \gamma_2)$, respectively, uniformly over $\gamma_1, \gamma_2 \in \Gamma$.*

Let $S_n(\gamma) = (1/\sqrt{n})\sum_{t=1}^n s_t(\gamma)$, and let $S(\gamma)$ denote a mean zero Gaussian process with covariance kernel $K(\gamma_1, \gamma_2)$. This means that for any $\{\gamma_1, \ldots, \gamma_k\} \in \Gamma^k$, $\{S(\gamma_1), \ldots, S(\gamma_k)\}$ is multivariate normal with mean zero and covariances $E(S(\gamma_j)S(\gamma_l)') = K(\gamma_j, \gamma_l)$. Next, let $\bar{S}(\gamma) = R'M(\gamma, \gamma)^{-1}S(\gamma)$ which is a mean-zero Gaussian process with covariance kernel

$$\bar{K}(\gamma_1, \gamma_2) = R'M(\gamma_1, \gamma_1)^{-1}K(\gamma_1, \gamma_2)M(\gamma_2, \gamma_2)^{-1}R.$$

THEOREM 1: *Under Assumptions* 1–3, $S_n \Rightarrow S$, $T_n \Rightarrow T^c$, *and* $g_n \Rightarrow g^c = g(T^c)$, *where*

$$T^c(\gamma) = (\bar{S}(\gamma)' + c'\bar{Q}(\gamma)')\bar{K}(\gamma,\gamma)^{-1}(\bar{S}(\gamma) + \bar{Q}(\gamma)c)$$

*and* $\bar{Q}(\gamma) = R'M(\gamma,\gamma)^{-1}M(\gamma,\gamma_0)R$, *with* $\gamma_0$ *the true value of* $\gamma$ *when* $c \neq 0$.

Theorem 1 gives the asymptotic distribution of the test statistic sequence $T_n$ and the global test $g_n$ under the local alternative $\beta_2 = c/\sqrt{n}$. Thus the asymptotic null distribution of $T_n$ is $T^0(\gamma) = \bar{S}(\gamma)'\bar{K}(\gamma,\gamma)^{-1}\bar{S}(\gamma)$, which has for each $\gamma \in \Gamma$ a marginal chi-square distribution. Hence, we call $T^0$ a *chi-square process*, and similarly $T^c$ a *noncentral chi-square process*. Since the null distribution of $g^0 = g(T^0)$ depends, in general, upon the covariance function $\bar{K}$, critical values cannot be tabulated except in special cases.[3]

## 3. THE *P*-VALUE TRANSFORMATION

Let $F^0(\cdot)$ denote the distribution function of $g^0$, and define $p_n = 1 - F^0(g_n)$. Tests based on $g_n$ and $p_n$ are equivalent since $F^0$ is monotonic and continuous. From Theorem 1 we see that $p_n \Rightarrow p^c$, where $p^c = 1 - F^0(g^c)$. In particular, the null distribution is $p^0 \sim U$, the uniform distribution on $[0,1]$. Thus the asymptotic null distribution of $p_n$ is free of nuisance parameters. Our test is to reject $H_0$ if $p_n \leq \alpha$. When the exact null distribution of $p_n$ is $U$, then $\alpha$ represents the size of the test; otherwise $\alpha$ is an asymptotic size, since

$$\lim_{n\to\infty} P\{p_n \leq \alpha | c = 0\} = \alpha. \tag{2}$$

We call $p_n$ an "asymptotic $p$-value."

Strictly speaking, the label "$p$-value" and "asymptotic size" are correct only if the convergence in (2) is uniform[4] over the null hypothesis. Note, however, that the distributions of $T_n$, $g_n$, and $p_n$ are invariant to the parameters $\beta_1$ and $\sigma^2$. Thus in the special case in which $x_t$ is strictly exogeneous and $\varepsilon_t$ is i.i.d. normal, the distribution of $p_n$ is free of nuisance parameters, so $H_0$ is simple and the convergence in (2) is trivially uniform. In the general regression context, uniform convergence may not hold. This is common in econometric testing problems and will not be emphasized in this paper.

The asymptotic power function associated with the test is

$$\pi_\alpha(c) = \lim_{n\to\infty} P\{p_n \leq \alpha | c\} = P\{F^0(g^c) \geq 1 - \alpha | c\}.$$

It is possible to show that as $|c|$ becomes large, $g^c \to \infty$ almost surely, so $p^c \to 0$ a.s. Hence for any $\alpha > 0$, $\pi_\alpha(c) \to 1$ as $|c| \to \infty$, and the test has nontrivial local power.

[3]For example, when testing for structural change of unknown timing (Andrews (1993b)) the asymptotic distribution only depends upon $\Gamma$ and $m$.

[4]For a definition see Sweeting (1980).

The random variable $g^0$ can be written as a continuous functional of the Gaussian process $S(\gamma)$, which is completely described by its covariance kernel $K(\gamma_1, \gamma_2)$. We can construct two estimates of this kernel, derived by the Wald and LM principles: $\hat{K}_n(\gamma_1, \gamma_2) = \frac{1}{n}\sum_{t=1}^n \hat{s}_t(\gamma_1)\hat{s}_t(\gamma_2)'$ and $\tilde{K}_n(\gamma_1, \gamma_2) = \frac{1}{n}\sum_{t=1}^n \tilde{s}_t(\gamma_1)\tilde{s}_t(\gamma_2)'$. One attraction of the estimator $\tilde{K}_n$ is that, like an LM statistic, it only involves estimation under the null hypothesis.

Now operate conditionally on the sample $w$. Denote by $\hat{S}_n$ a conditional mean-zero Gaussian process with covariance kernel $\hat{K}_n(\cdot,\cdot)$, by $\hat{T}_n$ a conditional chi-square process with covariance kernel $R'M_n(\gamma_1, \gamma_1)^{-1}\hat{K}_n(\gamma_1, \gamma_2)M_n(\gamma_2, \gamma_2)^{-1}R$, and let $\hat{g}_n = g(\hat{T}_n)$. Let $\hat{F}_n$ denote the conditional distribution function of $\hat{g}_n$, conditional on the sample $w$, and set $\hat{p}_n = 1 - \hat{F}_n(g_n)$. Similarly, let $\tilde{S}_n$ denote a conditional mean-zero Gaussian process with covariance kernel $\tilde{K}_n(\cdot,\cdot)$, and similarly $\tilde{T}_n$, $\tilde{g}_n$, $\tilde{F}_n$, and $\tilde{p}_n$.

These are not vacuous definitions. $\hat{S}_n$ and $\hat{T}_n$ can be generated by letting $\{v_t\}_{t=1}^n$ be i.i.d. $N(0,1)$ random variables, and setting

$$\hat{S}_n(\gamma) = \frac{1}{\sqrt{n}} \sum_{t=1}^n \hat{s}_t(\gamma)v_t \tag{3}$$

and

$$\hat{T}_n(\gamma) = \hat{S}_n(\gamma)'M_n(\gamma,\gamma)^{-1}R\left(R'\hat{V}_n^*(\gamma)R\right)^{-1}R'M_n(\gamma,\gamma)^{-1}\hat{S}_n(\gamma). \tag{4}$$

Similarly, $\tilde{S}_n$ and $\tilde{T}_n$ can be generated by

$$\hat{S}_n(\gamma) = \frac{1}{\sqrt{n}} \sum_{t=1}^n \tilde{s}_t(\gamma)v_t \tag{5}$$

and

$$\tilde{T}_n(\gamma) = \tilde{S}_n(\gamma)'M_n(\gamma,\gamma)^{-1}R\left(R'\hat{V}_n^*(\gamma)R\right)^{-1}R'M_n(\gamma,\gamma)^{-1}\tilde{S}_n(\gamma). \tag{6}$$

THEOREM 2: *Under Assumptions* 1 *and* 2, $\hat{p}_n = p_n + o_p(1)$ *and* $\tilde{p}_n = p_n + o_p(1)$. *Hence* $\hat{p}_n \Rightarrow p^c$ *and* $\tilde{p}_n \Rightarrow p^c$, *and under* $H_0$ *the asymptotic distribution of both* $\hat{p}_n$ *and* $\tilde{p}_n$ *is* $U$.

Theorem 2 shows that $\hat{p}_n$ and $\tilde{p}_n$ are asymptotically equivalent to $p_n$, under both the null hypothesis and local alternative $\beta_2 = c/\sqrt{n}$. The proof of the Theorem runs roughly as follows. We show that $\hat{S}_n \Rightarrow_p S$, where "$\Rightarrow_p$" denotes "weak convergence in probability" as defined by Gine and Zinn (1990). This implies that $\hat{g}_n \Rightarrow_p g^0$ and hence $\hat{F}_n(x) \to_p F^0(x)$, uniformly in $x$. Thus $\hat{p}_n = 1 - F^0(g_n) + o_p(1)$, as stated. The proof has similarities to those used to derive asymptotic properties of the bootstrap.

The conditional distribution functions $\hat{F}_n(\cdot)$ and $\tilde{F}_n(\cdot)$ are not directly observed, so neither are the random variables $\hat{p}_n$ and $\tilde{p}_n$. We can approximate either $\hat{F}_n$ or $\tilde{F}_n$ to any desired degree of accuracy, however, using standard

simulation techniques. Equations (3)–(6) show that a random draw from either conditional distribution can be made using a sample of iid $N(0,1)$ variables. Take, for example, $\hat{F}_n$. For $j = 1, \ldots, J$, execute the following steps:

(i) generate $\{v_{tj}\}_{t=1}^n$ iid $N(0,1)$ random variables (using a random number generator);

(ii) set $S_n^j(\gamma) = (1/\sqrt{n})\sum_{t=1}^n \hat{s}_t(\gamma)v_{tj}$;

(iii) set $T_n^j(\gamma) = S_n^j(\gamma)'M_n(\gamma,\gamma)^{-1}R(R'\hat{V}_n^*(\gamma)R)^{-1}R'M_n(\gamma,\gamma)^{-1}S_n^j(\gamma)$;

(iv) set $g_n^j = g(T_n^j)$.

This gives a random sample $(g_n^1, \ldots, g_n^J)$ of $J$ observations from the conditional distribution $\hat{F}_n$. Then compute the percentage of these artificial observations which exceed the actual test statistic $g_n$: $\hat{p}_n^J = \frac{1}{J}\sum_{j=1}^J \{g_n^j \geq g_n\}$. By the Glivenko-Cantelli Theorem, for any sample $\hat{p}_n^J \to_p 1 - \hat{F}_n(g_n) = \hat{p}_n$ as $J \to \infty$. Since $J$ is under the control of the econometrician, $\hat{p}_n^J$ can be made arbitrarily close to $\hat{p}_n$ by picking sufficiently large $J$, and thus can be used as our test statistic in place of $\hat{p}_n$. Analogously, we can construct $\tilde{p}_n^J$, using $S_n^j(\gamma) = (1/\sqrt{n})\sum_{t=1}^n x_t(\gamma)\tilde{\varepsilon}_t v_{tj}$ instead in step (ii). To select $J$ in practice, an appeal to the central limit theorem gives an asymptotic standard error for $\hat{p}_n^J$ of $\sqrt{\hat{p}_n^J(1-\hat{p}_n^J)/J}$. For example, when $\hat{p}_n^J = 0.05$, setting $J = 1000$ yields a standard error of only 0.007.

When the regression error is conditionally homoskedastic: $E(\varepsilon_t^2|\mathcal{F}_{t-1}) = \sigma^2$ a.s., the asymptotic expressions are significantly simpler. To approximate the asymptotic $p$-value using this information, for $j = 1, \ldots, J$, set $S_n^j(\gamma) = (1/\sqrt{n})\sum_{t=1}^n x_t(\gamma)v_{tj}$, where $v_{tj}$ are iid $N(0,1)$ draws from a random number generator,

$$T_n^j(\gamma) = S_n(\gamma)'M_n(\gamma,\gamma)^{-1}R\left(R'M_n(\gamma,\gamma)^{-1}R\right)^{-1}R'M_n(\gamma,\gamma)^{-1}S_n(\gamma),$$

and $\hat{p}_n^J = (1/J)\sum_{j=1}^J \{g(T_n^j) \geq g_n\}$. The same arguments as before show that $\hat{p}_n^J \to_p \hat{p}_n$ as $J \to \infty$ and $\hat{p}_n \to_d p^c$ as $n \to \infty$, with the null distribution $p^0 \sim U$.

If $\Gamma$ is a continuous parameter space, calculation of $g_n$ and replications $g_n^j$ might be excessively costly. In this case, it may be reasonable to replace $\Gamma$ by a discrete approximation $\Gamma_A = (\gamma_1, \ldots, \gamma_A)$. Then the actual test statistic is $g_{An} = g(T_n, \Gamma_A)$, where the notation is explicit about the dependence upon the region $\Gamma_A$. The simulated draws are $\hat{g}_{An}^j = g(\hat{T}_n^j, \Gamma_A)$ and the $p$-values are $\hat{p}_{An}^J = (1/J)\sum_{j=1}^J \{\hat{g}_{An}^j \geq g_{An}\}$. The null asymptotic theory goes through unaffected. Specifically, $g_{An} \Rightarrow g(T, \Gamma_A) = g_A$, say, and $\hat{p}_{An} \to^d p_A^c$, with $p_A^0 \sim U$. The power of the test may be adversely affected, however, if the selected approximation $\Gamma_A$ is insufficiently dense in $\Gamma$. Essentially, the function $T_n(\gamma)$ may have a sharp peak at some $\gamma_n \in \Gamma$ which is missed by the set $\Gamma_A$. A general solution to this problem may be impossible, since the appropriate choice of $\Gamma_A$ will depend upon the smoothness of $T_n(\gamma)$. If this is the case, several choices of $\Gamma_A$ could be used to assess sensitivity. In some applications, such as the threshold models explored in the next section, this is not an issue, as the sample function $T_n(\gamma)$ are naturally step functions with known step-points.

## 4. THRESHOLD REGRESSION

A typical threshold regression model takes the form (1) with $h(x_t, \gamma) = \{z_t \leq \gamma\} x_t$, where $\{\cdot\}$ denotes the indicator function, $z_t$ is an element of $x_t$, $x_{1t} = x_t$, and $\Gamma = [\gamma_L, \gamma_U]$. Threshold models are quite common in applied econometrics, typically interpreted as sample splits. By formally treating $\gamma$ as an unknown parameter, the threshold model allows the selection of $\gamma$ to be made conditional on the data.

The regression function and all test statistics are functions of $z_t$ and $\gamma$ only through the indicator function $\{z_t \leq \gamma\}$. For any monotonic transformation $Z_n(\cdot): \Gamma \to \Upsilon$, we have $\{z_t \leq \gamma\} = \{z_{nt} \leq \tau\}$ where $z_{nt} = Z_n(z_t)$ and $\tau = Z_n(\gamma)$. An ideal choice for $Z_n(\cdot)$ is the empirical distribution function of $z_t$. One advantage is that the transformed threshold parameter $\tau$ is unit-free and lies in the generic region [0, 1]. The pointwise test statistics are ill-behaved for extreme values of $\tau$, so $\Upsilon$ should not include values of $\tau$ too close to 0 or 1. There is no obvious criteria by which to make this choice, but a similar dilemma appears in the changepoint literature, where recently Andrews (1993b) suggested $\Upsilon = [.15, .85]$.

Davies (1977) mentioned threshold models as a possible application of his distributional theory, but did not investigate whether his conditions were satisfied. In fact, they are not. Davies assumed that the limit process $T(\gamma)$ has a derivative, but this is not the case in threshold models. Take the special case where $x_t = 1$ and $z_t = y_{t-1}$. Here, the asymptotic process $T(\gamma)$ is known to be a normalized squared Brownian bridge (see Chan (1990)), which is nowhere differentiable and thus has infinite total variation. The bound Davies uses is a function of the sample total variation, which is finite in any sample yet diverges to infinity as $n \to \infty$, rendering an asymptotic size of zero. Table I compares Davies' procedure with the test based on the asymptotic critical values, where the innovations $\varepsilon_t$ are iid $N(0, 1)$, and the sample size is varied from 50 to 1000. The null sets $\theta = 0$ and the alternative sets $\theta = .5$ and $\gamma = 0$. The first five columns show the rejection frequency under the null, in which case Davies' suggested procedure is extremely conservative. The cost shows up in the rejection frequency under the alternative (the final five columns), where the Davies' procedure yields tests with power less than the nominal 5% rate for samples of 500 or less. As expected from the asymptotic theory, Davies' bound is not useful in this context.

TABLE I

SIZE AND POWER OF NOMINAL 5% SIZE TESTS FOR SHIFT IN MEAN

| | Null $\theta = 0$ | | | | | Alternative $\theta = .5$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sample Size | 50 | 100 | 250 | 500 | 1000 | 50 | 100 | 250 | 500 | 1000 |
| Davies Procedure | .00 | .01 | .01 | .01 | .00 | .01 | .02 | .03 | .05 | .09 |
| Asymptotic Criticals | .02 | .04 | .04 | .04 | .05 | .04 | .08 | .12 | .19 | .34 |

Instead, we can apply the conditional transformation of Section 3 to obtain appropriate asymptotic critical values. The following theorem shows that this procedure is justified.

THEOREM 3: *Assume that Assumption* 1 *holds with* $r > v$, *and* $z_t$ *has density function* $f(z)$ *such that* $\sup_{x \in R} f(x) = \bar{f} < \infty$. *Then Assumptions* 2 *and* 3 *are satisfied, and hence Theorems* 1 *and* 2 *hold.*

The threshold model assumes that the threshold variable $z_t$ is known a priori. This is not always the case. In many examples, all that is known is that $z_t$ is some element of $x_{2t}$, some $k_2 \times 1$ subvector of $x_t$. In this case, we can write $z_t = x_{2t}(d)$, the $d$th element of $x_{2t}$. In this notation, we can think of $d$ as a parameter, taking values in the discrete set $D = \{1, 2, \ldots, k_2\}$. The standard threshold model emerges as the special case in which $d$ is known a priori. In the threshold selection model $d$ is unknown and must be estimated along with the other parameters.

For fixed $(\gamma, d)$, the model is linear in $(\alpha, \theta)$ so can be estimated by ordinary least squares. The global estimates of the parameters can be found by minimization of the resulting least squares variance estimate over $(\gamma, d) \in (\Gamma \times D)$. Under the null hypothesis both the parameters $\gamma$ and $d$ are not identified, so the pair $(\gamma, d)$ is treated as we had treated $\gamma$ in the earlier sections. Thus, the pointwise test statistics $T_n(\gamma, d)$ may be found for each $(\gamma, d)$, and the transformations $g_n$, $\hat{p}_n$, and $\hat{p}_n^J$ found as before, replacing the argument $\gamma$ by $(\gamma, d)$.

It is not hard to see that the asymptotic theory is essentially unaffected. Since $D$ is a finite set, all convergence results are uniform over $d \in D$.

COROLLARY 1: *If* $z_t = x_{2t}(d)$, *assume that Assumption* 1 *holds with* $r > v$, *and* $x_{2t}$ *has density function* $f(x)$ *such that* $\sup_{x \in R^{k_2}} f(x) = \bar{f} < \infty$. *Then Assumptions* 2 *and* 3 *are satisfied, and hence Theorems* 1 *and* 2 *hold for the threshold selection model.*

A special example of a threshold regression is the so-called *self-exciting threshold autoregressive model* (SETAR), which has received considerable recent attention in the nonlinear time series literature. The model takes the form

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + (\theta_0 + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p})\{y_{t-d} \le \gamma\} + \varepsilon_t \tag{7}$$

with $\{\varepsilon_t, \mathscr{F}_t\}$ a MDS. The model (7) falls in the class of threshold selection models by setting $x_t = (1 \;\; y_{t-1} \;\; \cdots \;\; y_{t-p})$ and $z_t(d) = y_{t-d}$, and is largely due to Tong (1983). Chan (1990) proved an analog of Theorem 1 under the assumption that $d$ is known and $\varepsilon_t$ is iid Gaussian. Chan (1991) approximated the tail of the asymptotic distribution for general $p$ (but still fixed and known $d$). General results, allowing arbitrary distribution functions for $\varepsilon_t$, general functionals on $\Gamma$ such as the Andrews-Ploberger optimal transformations, and treating $d$ as an

unknown parameter, have not been treated before. Our theory easily handles these generalizations.

COROLLARY 2: *Suppose that* $\{y_t\}$ *is generated by* (7) *with* $\varepsilon_t$ *independent and identically distributed, all the roots of the characteristic equation* $z^p - \alpha_1 z^{p-1} - \cdots - \alpha_p = 0$ *lie strictly within the unit circle,* $E|\varepsilon_t|^{4r} < \infty$ *for some* $r > 1$, *and the density of* $\varepsilon_t$ *is bounded and continuous. Then under* $H_0: \theta_0 = \cdots = \theta_p = 0$, $\hat{p}_n \Rightarrow U$ *and* $\tilde{p}_n \Rightarrow U$.

The assumptions of Corollary 2 imply that $y_t$ is strictly stationary and absolutely regular with exponentially declining mixing coefficients (see Pham and Tran (1985)), establishing Assumption 1. It also implies that $y_t$ has a bounded density, completing the requirements for Corollary 1. The only reason why we technically restrict attention to the null hypothesis is because it has not been investigated whether or not SETAR processes are absolutely regular. Similarly, the restriction that $\varepsilon_t$ is iid is made only to guarantee absolute regularity. It is quite likely that this condition is not necessary and could be replaced by a martingale difference condition. Hence while the assumptions include homoskedasticity, we do not think that it is prudent to impose this condition when constructing test statistics.

## 5. FINITE SAMPLE DISTRIBUTION

To assess the usefulness of the testing methodology, finite sample distributional results are reported in a simple Monte Carlo simulation study. We use the threshold-selection SETAR model (7) with $p = 1$, 2, and 3, $\alpha_1 = .3$, $\alpha_0 = \alpha_2 = \alpha_3 = 0$, $\varepsilon_t$ iid $N(0,1)$ and two sample sizes, $n = 100$ and $n = 200$. When $p > 1$, both the threshold lag $d$ and the threshold $\gamma$ were estimated by least squares. The experiments were done for tests of size 10%, 5%, and 1%, but only those for size 5% are reported since no differences were observed.

The test statistics functions $T_n(\gamma, d)$ were calculated using four different covariance matrices: (i) standard Wald (W); (ii) standard Lagrange multiplier (LM); (iii) MacKinnon-White (1985) "jackknife" heteroskedasticity-consistent Wals ($W^h$); (iv) Eicker-White heteroskedasticity-consistent Lagrange multiplier ($LM^h$). From each of these test statistic functions, two functionals were used: the supremum (sup), and average (ave) (see Section 2) to generate the test statistics $g_n$ and the simulated $p$-values $\hat{p}_n^J$. Due to the large computational requirements of the simulation design, the number of internal simulation replications was set at $J = 500$ and the region $\Upsilon$ for the calculation of the test statistics (see Section 4) was set at $\Upsilon = [.2, .8]$. A ninth test statistic, included for comparison, is the $S_3$ test of Luukkonen, Saikkonen, and Terasvirta (1988), which is an LM-type test for a smooth transition autoregressive (STAR) model. The simulation study reported in Luukkonen, et al., showed that the $S_3$ test has good power.

TABLE II

FINITE SAMPLE SIZE OF ASYMPTOTIC 5% TESTS

| | $n = 100$ | | | $n = 200$ | | |
|---|---|---|---|---|---|---|
| AR order | $p = 1$ | $p = 2$ | $p = 3$ | $p = 1$ | $p = 2$ | $p = 3$ |
| SupW | .07 | .07 | .07 | .05 | .05 | .08 |
| AveW | .05 | .05 | .04 | .04 | .04 | .06 |
| SupLM | .05 | .04 | .03 | .04 | .04 | .05 |
| AveLM | .05 | .04 | .03 | .04 | .04 | .06 |
| $\text{SupW}^h$ | .14 | .21 | .32 | .09 | .13 | .19 |
| $\text{AveW}^h$ | .07 | .09 | .10 | .05 | .06 | .07 |
| $\text{SupLM}^h$ | .04 | .02 | .03 | .05 | .04 | .02 |
| $\text{AveLM}^h$ | .04 | .04 | .03 | .04 | .04 | .05 |
| STAR | .05 | .04 | .03 | .04 | .04 | .04 |

We report in Table II the actual size (setting $\theta_0 = \theta_1 = 0$). For samples of size 100, reported in the first three columns, 2000 simulated samples were drawn. For samples of size 200, reported in the final three columns, 1000 simulated samples were used for $p = 1$ and $p = 2$, and 500 for $p = 3$. We find that the asymptotic approximation is excellent for the "standard" tests (without the heteroskedasticity correction). The heteroskedasticity-consistent LM tests are slightly conservative for large $p$, and the heteroskedasticity-consistent Wald statistic $\text{supW}^h$ is excessively liberal, especially at $n = 100$. Based on these results, our recommendation is to avoid the $\text{supW}^h$ statistic unless the sample size is very large.

To assess power, we consider two specifications, using first an intercept shift ($\theta_0 \neq 0$) and second a slope shift ($\theta_1 \neq 0$). In both specifications, $\gamma$ is selected so that it is approximately the median of $y_{t-1}$. (Since there is no closed-form expression for the distribution or density of $y_t$, the median was calculated by simulation.) For each simulation design with $p \leq 2$, 1000 replications were made, and 500 replications were made when $p = 3$. Finite sample 5% critical values for the test statistics were calculated from the .05 percentile of the empirical distribution of the tests calculated under the null. For each $p$ and each $n$, the power of the Wald and LM tests were nearly identical, so we only report in Table III the results of the LM-based tests.

As expected, power is increasing in $|\theta_0|$ or $|\theta_1|$, increasing in $n$, and decreasing with $p$. Indeed, the effect of the estimated model order ($p$) is quite strong, indicating that the cost of over-fitting is high. The theory of Andrews-Ploberger (1994) suggests that aveLM should be optimal against local alternatives, yet this is not supported by the simulation evidence. Instead, we find that for shifts in the intercept supLM has the best power, yet for shifts in the slope aveLM dominates. It is also interesting to compare the power of our tests with the STAR test of Luukkonen, et al. The STAR test does remarkably well against the shift in slope, although less well against the shift in the intercept. It also does better against local alternatives than distant alternatives, which is expected since it was derived using a local power argument. It is also possible to see a slight

TABLE III

FINITE SAMPLE POWER OF 5% SIZE TESTS

| | | $n = 100$ | | | | $n = 200$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\theta_0$ | | $\theta_1$ | | $\theta_0$ | | $\theta_1$ | |
| | | −.75 | −1.25 | −.6 | −1.0 | −.75 | −1.25 | −.6 | −1.0 |
| $p = 1$ | | | | | | | | | |
| | SupLM | .29 | .70 | .25 | .69 | .63 | .99 | .55 | .98 |
| | SupLM$^h$ | .28 | .65 | .22 | .62 | .61 | .98 | .52 | .96 |
| | AveLM | .21 | .50 | .38 | .84 | .39 | .86 | .67 | .99 |
| | AveLM$^h$ | .20 | .47 | .36 | .80 | .35 | .83 | .63 | .98 |
| | STAR | .18 | .42 | .31 | .73 | .34 | .72 | .58 | .97 |
| $p = 2$ | | | | | | | | | |
| | SupLM | .22 | .61 | .18 | .60 | .29 | .70 | .25 | .69 |
| | SupLM$^h$ | .20 | .55 | .16 | .52 | .28 | .65 | .22 | .62 |
| | AveLM | .13 | .27 | .25 | .67 | .21 | .50 | .38 | .84 |
| | AveLM$^h$ | .12 | .26 | .21 | .61 | .20 | .47 | .35 | .80 |
| | STAR | .14 | .30 | .19 | .56 | .18 | .42 | .31 | .73 |
| $p = 3$ | | | | | | | | | |
| | supLM | .15 | .48 | .18 | .52 | .15 | .48 | .18 | .52 |
| | SupLM$^h$ | .15 | .44 | .18 | .48 | .15 | .44 | .18 | .48 |
| | AveLM | .08 | .17 | .19 | .54 | .08 | .17 | .19 | .54 |
| | AveLM$^h$ | .07 | .16 | .18 | .47 | .07 | .16 | .18 | .47 |
| | STAR | .10 | .21 | .15 | .43 | .10 | .21 | .15 | .43 |

deterioration in power from the use of the heteroskedasticity-consistent test statistics, rather than the tests which assume homoskedasticity. The power loss is fairly mild, however. In sum, the simulation evidence strongly favors using our new tests.

## 6. APPLICATION TO U.S. GNP

We now apply this testing methodology to Potter's (1995) model of U.S. GNP. He used a SETAR in growth rates with a first, second, and fifth autoregressive lag. To select the threshold and delay parameters, he used informal graphical methods. While not an optimal estimation method, this is still conditional on the data, and hence invalidates the use of conventional test statistics and critical values. Our tests, on the other hand, allow a direct assessment of the statistical significance of his model.

The data is real GNP (seasonally adjusted) from Citibase for the period 1947–1990 and transformed into annualized quarterly growth rates. We estimated the model by least squares, allowing the threshold variable $\gamma$ to vary from the 15th to the 85th percentile of the empirical distribution of $x_t$, and the delay parameter $d$ over 1, 2, and 5. Our estimates are $\hat{d} = 2$, $\hat{\gamma} = 0.01$, and:

Regime 1—$x_{t-2} \le 0.01$:

$$x_t = \underset{(1.12)}{-3.21} + \underset{(.25)}{.51}\, x_{t-1} - \underset{(.31)}{.93}\, x_{t-2} - \underset{(.25)}{.38}\, x_{t-5} + \hat{\varepsilon}_t, \quad \hat{\sigma}_1^2 = 23.5;$$

TABLE IV

TESTS FOR THRESHOLD EFFECT IN U.S. GNP: 1947–1990

| | SupLM | ExpLM | AveLM | SupLM$^h$ | ExpLM$^h$ | AveLM$^h$ |
|---|---|---|---|---|---|---|
| $g_n$ | 18.2 | 4.8 | 4.6 | 14.1 | 4.0 | 4.7 |
| $\hat{p}_n^J$ | 0.04 | 0.09 | 0.29 | 0.17 | 0.17 | 0.27 |

Regime 2—$x_{t-2} > 0.01$:

$$x_t = \underset{(.77)}{2.14} + \underset{(.10)}{.30}\, x_{t-1} + \underset{(.10)}{.18}\, x_{t-2} - \underset{(.07)}{.16}\, x_{t-5} + \hat{\varepsilon}_t, \quad \hat{\sigma}_2^2 - 12.1.$$

Heteroskedastic-consistent standard errors are given in parenthesis.

The LM-based tests for the hypothesis of a single regime (no threshold effect) are reported in Table IV, with mixed results. If the homoskedasticity hypothesis is maintained, then supLM is marginally significant at the asymptotic 5% level, while aveLM is far from the rejection region. The point estimates for the error variance in the two regimes suggest that there may indeed be error heteroskedasticity. The test statistics which are robust to heteroskedasticity, however, are all far from standard rejection regions. Should we believe the "rejection" implied by supLM or the caution implied by the other statistics? The marginally significant supLM statistic could possibly be due to heteroskedasticity, yet the insignificant supLM$^h$ statistic could possibly be due to its lower power. Our simulation evidence, however, suggested that the power loss is small. No definitive answer is possible, but it is fair to conclude that the apparent significance of the threshold model is consistent with sampling variation.

## 7. CONCLUSION

Many econometric models contain unidentified parameters under interesting parametric restrictions. Tests concerning these restrictions cannot use conventional statistical theory to assess significance. This paper extends the literature by developing a simple simulation technique which produces $p$-value statistics with an asymptotic uniform distribution under the null hypothesis.

*Dept. of Economics, Boston College, Chestnut Hill, MA 02167, U.S.A.*

*Manuscript received September, 1991; final revision received May, 1995.*

## APPENDIX: MATHEMATICAL PROOFS

PROOF OF THEOREM 1: We start by showing that $S_n \Rightarrow S$. Note that for each $\gamma \in \Gamma$, $s_t(\gamma) = x_t(\gamma)\varepsilon_t$ is a square integrable stationary martingale difference, to which the pointwise central limit

theorem applies. Furthermore, note that the covariance kernel for $S_n(\gamma)$ is

$$E(S_n(\gamma_1)S_n(\gamma_2)') = \frac{1}{n}\sum_{t=1}^{n} E(s_t(\gamma_1)s_t(\gamma_2)') = K(\gamma_1,\gamma_2).$$

The multivariate central limit theorem establishes the finite dimensional distributional convergence.

To establish stochastic equicontinuity, we appeal to Theorem 1, Application 5, of Doukhan, Massart, and Rio (1994). Our data, and hence the summands $s_t(\gamma)$, satisfy the necessary absolute regularity mixing decay rate, and the envelope function is $\sup_\gamma |s_t(\gamma)| = \bar{x}_t|\varepsilon_t|$, which is $L^{2v}$ bounded under Assumption 1. Finally, we need to show that the log of the $L^{2v}$ bracketing numbers is integrable. Since $\Gamma \subset R^q$, one can always find a set $\Gamma_N$ and constant $G < \infty$ so that for all $\gamma$ there is some $\gamma_k \in \Gamma_N$ satisfying

$$|\gamma - \gamma_k| \le GN^{-1/q}. \tag{8}$$

Set $N(\delta) = G^q B^{q/\lambda}\delta^{q/\lambda}$. Using Assumption 2 and (8), for all $\gamma \in \Gamma$,

$$\begin{aligned} \|s_t(\gamma) - s(\gamma_k)\|_{2v} &= \|(h_t(\gamma) - h_t(\gamma_k))\varepsilon_t\|_{2v} \\ &\le B|\gamma - \gamma_k|^\lambda \le BG^\lambda N^{-\lambda/q} = \delta. \end{aligned} \tag{9}$$

Thus $N(\delta)$ satisfy the definition of the $L^{2v}$ bracketing numbers. Since

$$\int_0^1 \sqrt{\log(N(\delta))}\, d\delta \le \sqrt{q\log(GB^{1/\lambda})} + \int_0^1 \sqrt{\log(\delta^{-q/\lambda})}\, d\delta < \infty$$

the conditions of Doukhan, Massart, and Rio are met, establishing that $S_n$ is stochastically equicontinuous, and hence that $S_n \Rightarrow S$.

Standard algebra and Assumptions 1–3 show that

$$\begin{aligned} \sqrt{n}\, R'\hat{\beta}(\gamma) &= R'M_n(\gamma,\gamma)^{-1}S_n(\gamma) + R'M_n(\gamma,\gamma)^{-1}M_n(\gamma,\gamma_0)Rc \\ &\Rightarrow R'M(\gamma,\gamma)^{-1}S(\gamma) + R'M(\gamma,\gamma)^{-1}M(\gamma,\gamma_0)Rc = \bar{S}(\gamma) + \bar{Q}(\gamma)c. \end{aligned} \tag{10}$$

It is not hard to show that $\hat{V}_n^*(\gamma) \to K(\gamma,\gamma)$ a.s., uniformly over $\gamma$. Hence

$$R'\hat{V}_n^*(\gamma)R \to R'M(\gamma,\gamma)^{-1}K(\gamma,\gamma)M(\gamma,\gamma)^{-1}R = \bar{K}(\gamma,\gamma). \tag{11}$$

The continuous mapping theorem (CMT) applied to (10) and (11) show that $T_n \Rightarrow T^c$, and hence $g_n \Rightarrow g^c$ by the CMT. *Q.E.D.*

PROOF OF THEOREM 2: Let $s_t^*(\gamma) = s_t(\gamma)v_t$ and $S_n^*(\gamma) = (1/\sqrt{n})\Sigma_{t=}^n s_t^*(\gamma)$. We first show that $S_n^w \Rightarrow_p S$. Let $W$ denote the set of samples $w$ for which

$$\begin{aligned} &\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \bar{x}_t^2\varepsilon_t^2 < \infty, \\ &K_n(\gamma_1,\gamma_2) \to K(\gamma_1,\gamma_2), \qquad \text{uniformly over } \gamma_1 \text{ and } \gamma_2. \end{aligned} \tag{12}$$

Under Assumptions 1–3, $P(W) = 1$. Take any $w \in W$. For the remainder of the proof, we will be operating conditionally on $w$, so all of the randomness appears in the iid $N(0,1)$ variables $v_t$. Set $E_w X = E(X|w)$.

Note that $s_t^*(\gamma)$ are independent mean-zero normal random vectors, and $S_n^*$ is a mean-zero Gaussian process with covariance function

$$E_w(S_n^*(\gamma_1)S_n^*(\gamma_2)') = \frac{1}{n}\sum_{t=1}^{n} E_w(s_t(\gamma_1)s_t(\gamma_2)'v_t^2)$$

$$= \frac{1}{n}\sum_{t=1}^{n} s_t(\gamma_1)s_t(\gamma_2)' = K_n(\gamma_1, \gamma_2).$$

Since $K_n(\gamma_1, \gamma_2) \to K(\gamma_1, \gamma_2)$ by (12), the finite dimensional distribution of $S_n^*$ converge to those of $S$.

Next, the envelope of $s_t(\gamma)$ is $L^2$-integrable:

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} E_w \sup_{\gamma\in\Gamma} |s_t(\gamma)|^2 = \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \bar{x}_t^2 \varepsilon_t^2 < \infty,$$

by (12). Define the norms

$$\rho_n(\gamma_1, \gamma_2) = (E_w |S_n^*(\gamma_1) - S_n^*(\gamma_2)|^2)^{1/2}$$

$$= \left(\frac{1}{n}\sum_{t=1}^{n} (s_t(\gamma_1) - s_t(\gamma_2))^2\right)^{1/2}$$

and $\rho(\gamma_1, \gamma_2) = \|s_t(\gamma_1) - s_t(\gamma_2)\|_2$. Observe that

$$\rho_n(\gamma_1, \gamma_2) = \left[\operatorname{tr}\left\{\frac{1}{n}\sum_{t=1}^{n} (s_t(\gamma_1) - s_t(\gamma_2))(s_t(\gamma_1) - s_t(\gamma_2))'\right\}\right]^{1/2}$$

$$= [\operatorname{tr}\{K_n(\gamma_1, \gamma_1) + K_n(\gamma_2, \gamma_2) - K_n(\gamma_2, \gamma_1) - K_n(\gamma_1, \gamma_2)\}]^{1/2}$$

$$\to [\operatorname{tr}\{K(\gamma_1, \gamma_1) + K(\gamma_2, \gamma_2) - K(\gamma_2, \gamma_1) - K(\gamma_1, \gamma_2)\}]^{1/2} = \rho(\gamma_1, \gamma_2)$$

uniformly over $\gamma_1$ and $\gamma_2$ under (12).

For any integer $N$, let $\Gamma_N$ be the set satisfying (8), and for any $\delta > 0$ set $N(\delta) = G^q B^{q/\lambda} \delta^{q/\lambda}$. By the monotonicity of the $L^p$-norms and (9), we have

$$\rho(\gamma, \gamma_k) \le \|s_t(\gamma) - s_t(\gamma_k)\|_{2v} \le \delta,$$

establishing that $N(\delta)$ are the $L^2$ bracketing numbers, and have an exponential decay rate. Pollard (1990, Theorem 10.6) showed that $S_n^*$ is stochastically equicontinuous under these conditions. Hence (for the particular $w$ we selected), $S_n^w \Rightarrow S$. Since $P(W) = 1$, $S_n^* \Rightarrow S$, as desired.

Second, we wish to show that $\hat{S}_n \Rightarrow_p S$. Note that $\hat{S}_n(\gamma) = S_n^*(\gamma) + S_n^r(\gamma)$, where $S_n^r(\gamma) = (1/\sqrt{n})\sum_{t=1}^{n} x_t(\gamma)(\hat{\varepsilon}_t(\gamma) - \varepsilon_t)v_t$. We have

$$\sup_{\gamma\in\Gamma} |S_n^r(\gamma)| \le \sup_{\gamma\in\Gamma} \left|\frac{1}{n}\sum_{t=1}^{n} x_t(\gamma)x_t(\gamma)'v_t\right| \sup_{\gamma\in\Gamma} |\sqrt{n}(\hat{\beta}(\gamma) - \beta_0)|. \tag{13}$$

Using the same arguments as before, we can see that conditional on $w$, $(1/n)\sum_{t=1}^{n} x_t(\gamma)x_t(\gamma)'v_t \Rightarrow 0$ a.s. Thus (13) converges weakly in probability to zero, and $\hat{S}_n \Rightarrow_p S$.

This allows us to find under Assumptions 1–3 and the CMT that

$$\hat{T}_n(\gamma) = \hat{S}_n(\gamma)' M_n(\gamma, \gamma)^{-1} R(R'\hat{V}_n^*(\gamma)R)^{-1} R' M_n(\gamma, \gamma)^{-1} \hat{S}_n(\gamma)$$

$$\Rightarrow_p \bar{S}(\gamma)' \bar{K}(\gamma, \gamma)^{-1} \bar{S}(\gamma) = T^0(\gamma).$$

Thus $\hat{g}_n = g(\hat{T}_n) \Rightarrow_p g(T^0) = g^0$, which implies that $\hat{F}_n(x) \to_p F^0(x)$, uniformly in $x$. We conclude that

$$\hat{p}_n = 1 - \hat{F}_n(g_n) = 1 - F^0(g_n) + o_p(1) = p_n + o_p(1),$$

as stated. The proof for $\hat{p}_n$ is similar. *Q.E.D.*

The proof of Theorem 3 relies on the following uniform strong law.

LEMMA 1: *If* $\{w_t\}$ *is strictly stationary and ergodic,* $E|\phi(w_t)| < \infty$, *and* $w_t$ *has a continuous distribution, then*

$$\sup_{\gamma \in R} \left| \frac{1}{n} \sum_{t=1}^{n} \phi(w_t)\{w_t \le \gamma\} - E(\phi(w_t)\{w_t \le \gamma\}) \right| \to 0 \quad a.s. \tag{14}$$

PROOF OF LEMMA 1: Pollard (1984, Theorem II.2) established (14) when for each $\varepsilon > 0$ there exists a class of approximating functions $\{f^u_{\varepsilon,k}(w_t), f^l_{\varepsilon,k}(w_t) : k = 1, \ldots, K_\varepsilon\}$, $K_\varepsilon < \infty$, which have the property that for each $\gamma$, there exists some $k$ such that $f^l_{\varepsilon,k}(w_t) \le \phi(w_t)\{w_t \le \gamma\} \le f^u_{\varepsilon,k}(w_t)$ and $E|f^u_{\varepsilon,k}(w_t) - f^l_{\varepsilon,k}(w_t)| < \varepsilon$. We can construct a qualifying set of approximating functions as follows. Set $K_\varepsilon = 2E|\phi(w_t)|/\varepsilon$ and select $\{\gamma_1, \ldots, \gamma_{K_\varepsilon}\}$ so that for all $k$

$$E(|\phi(w_t)|\{\gamma_k < w_t \le \gamma_{k+1}\}) \le \varepsilon/2. \tag{15}$$

(This is possible since $w_t$ has a continuous distribution.) Then set

$$f^u_{\varepsilon,k}(w_t) = \phi(w_t)\{w_t \le \gamma_{k+1}\}\{\phi(w_t) \ge 0\} + \phi(w_t)\{w_t \le \gamma_k\}\{\phi(w_t) < 0\}$$

and

$$f^l_{\varepsilon,k}(w_t) = \phi(w_t)\{w_t \le \gamma_k\}\{\phi(w_t) \ge 0\} + \phi(w_t)\{w_t \le \gamma_{k+1}\}\{\phi(w_t) < 0\}.$$

By construction, for all $\gamma$, there is some $k$ such that $\gamma_k < \gamma \le \gamma_{k+1}$ and $f^l_{\varepsilon,k}(w_t) \le \phi(w_t)\{w_t \le \gamma\} \le f^u_{\varepsilon,k}(w_t)$. Hence,

$$\begin{aligned} E|f^u_{\varepsilon,k}(w_t) - f^l_{\varepsilon,k}(w_t)| &\le 2E|\phi(w_t)(\{w_t \le \gamma_{k+1}\} - \{w_t \le \gamma_k\})| \\ &= 2E(|\phi(w_t)|\{\gamma_k < w_t \le \gamma_{k+1}\}) \le \varepsilon, \end{aligned}$$

where the final inequality is (15). *Q.E.D.*

PROOF OF THEOREM 3: When $\gamma' < \gamma$, by Holder's inequality and the boundedness of $f$,

$$\begin{aligned} \|(h_t(\gamma) - h_t(\gamma'))\varepsilon_t\|_{2v} &= \|x_t \varepsilon_t \{\gamma' < z_t \le \gamma\}\|_{2v} \\ &\le \|x_t \varepsilon_t\|_{2r} \|\{\gamma' < z_t \le \gamma\}\|_{2rv/(r-v)} \\ &\le \|x_t\|_{4r} \|\varepsilon_t\|_{4r} \left( \int_{\gamma'}^{\gamma} f(x)\,dx \right)^{(r-v)/2rv} \le B|\gamma - \gamma'|^{\lambda} \end{aligned}$$

where $B = \|x_t\|_{4r}\|\varepsilon_t\|_{4r} \bar{f}^{(r-v)/2rv}$ and $\lambda = (r-v)/2rv$. The same inequality holds for $\gamma < \gamma'$, establishing Assumption 2.

Note that

$$M_n(\gamma_1, \gamma_2) = \begin{bmatrix} \frac{1}{n}\sum_{t=1}^{n} x_t x_t' & \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \{z_t \le \gamma_2\} \\ \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \{z_t \le \gamma_1\} & \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \{z_t \le \min(\gamma_1, \gamma_2)\} \end{bmatrix} \to M(\gamma_1, \gamma_2)$$

and

$$K_n(\gamma_1,\gamma_2)=\begin{bmatrix} \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \varepsilon_t^2 & \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \varepsilon_t^2\{z_t \le \gamma_2\} \\ \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \varepsilon_t^2\{z_t \le \gamma_1\} & \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \varepsilon_t^2\{z_t \le \min(\gamma_1,\gamma_2)\} \end{bmatrix} \to K(\gamma_1,\gamma_2)$$

a.s., uniformly in $\gamma_1$ and $\gamma_2$ by Lemma 1, and the assumption that $z_t$ has a continuous distribution. This establishes Assumption 3. *Q.E.D.*

## REFERENCES


ANDREWS, D. W. K. (1993a): "An Introduction to Econometric Applications of Functional Limit Theory for Dependent Random Variables," *Econometric Reviews*, 12, 183–216.

——— (1993b): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821–856.

ANDREWS, D. W. K., AND W. PLOBERGER (1994): "Optimal Tests when a Nuisance Parameter is Present Only under the Alternative," *Econometrica*, 62, 1383–1414.

BIERENS, J. H. (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1443–1458.

CHAN, K. S. (1990): "Testing for Threshold Autoregression," *The Annals of Statistics*, 18, 1886–1894.

——— (1991): "Percentage Points of Likelihood Ratio Tests for Threshold Autoregression," *Journal of the Royal Statistical Society*, Series B, 53, 691–696.

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

——— (1987): "Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 74, 33–43.

DOUKHAN, P., P. MASSART, AND E. RIO (1994): "Invariance Principles for Absolutely Regular Empirical Processes," *Annales de l'Institut H. Poincare*, forthcoming.

GINE, E., AND J. ZINN (1990): "Bootstrapping General Empirical Measures," *The Annals of Probability*, 18, 851–869.

HANSEN, B. E. (1991): "Inference when a Nuisance Parameter is not Identified under the Null Hypothesis," Rochester Center for Economic Research Working Paper No. 296.

KING, M. L., AND T. S. SHIVELY (1991): "Locally Optimal Testing when a Nuisance Parameter is Present only under the Alternative," unpublished manuscript, Monash University.

LUUKKONEN, R., P. SAIKKONEN, AND T. TERASVIRTA (1988): "Testing Linearity Against Smooth Transition Autoregressive Models," *Biometrika*, 75, 491–499.

MACKINNON, J. G., AND H. WHITE (1985): "Some Heteroskedasticity-consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325.

PHAM, T. D., AND L. T. TRAN (1985): "Some Mixing Properties of Time Series Models," *Stochastic Processes and their Applications*, 19, 297–303.

POLLARD, D. (1984): *Convergence of Stochastic Processes*. NY: Springer-Verlag.

——— (1990): *Empirical Processes: Theory and Applications*. Hayward: Institute of Mathematical Statistics.

POTTER, S. M. (1995): A Nonlinear Approach to U.S. GNP," *Journal of Applied Econometrics*, forthcoming.

STINCHCOMB, M. B., AND H. WHITE (1993): "Consistent Specification Testing with Unidentified Nuisance Parameters Using Duality and Banach Space Limit Theory," unpublished manuscript, UCSD.

SWEETING, T. J. (1980): "Uniform Asymptotic Normality of the Maximum Likelihood Estimator," *The Annals of Statistics*, 8, 1375–1381.

TONG, H. (1983): *Threshold Models in Non-linear Time Series Analysis*, Lecture Notes in Statistics, 21. Berlin: Springer.

VOLKONSKII, V. A., AND Y. A. ROZANOV (1959): "Some Limit Theorems for Random Functions, Part I," *Theory of Probability and Its Application*, 4, 178–197.

WHITE, H. (1989): "An Additional Hidden Unit Test for Neglected Nonlinearity in Multilayer Feedforward Networks," *Proceeding of the International Joint Conference on Neural Networks*, II, Washington, D.C. New York, NY: IEEE Press, 451–455.