The Likelihood Test Under Non-Standard Conditions: Testing the Markov Trend Model of GDP

Bruce E. Hansen

Working Paper No. 279
May 1991

The Likelihood Test Under Non-Standard Conditions: Testing the Markov Trend Model of GNP

Bruce E. Hansen

Rochester Center for Economic Research
Working Paper No. 279

May 1991

The Likelihood Test Under Non–Standard Conditions:
Testing The Markov Trend Model of GNP

Bruce E. Hansen

University of Rochester

and

Rochester Center for Economic Research

Working Paper No. 279

March 1991

I would like to thank Fabio Canova, Adrian Pagan and Changyong Rhee for helpful conversations during this project, James Hamilton for helpful comments and kindly providing his data and computer programs, and the NSF for financial support under grant # SES 9022176.

Abstract

A theory of testing under non–standard conditions is developed. By viewing the likelihood as a function of the unknown parameters, empirical process theory enables us to bound the asymptotic distribution of likelihood ratio statistics, even when conventional regularity conditions (such as unidentified nuisance parameters and identically zero scores) are violated. This testing methodology is applied to the Markov trend model of GNP proposed by Hamilton (1989). The standardized likelihood ratio test is unable to reject the hypothesis of an AR(4) in favor of the Markov trend model. Instead, we find strong evidence for an alternative model, which we call a *mixture model* of GNP, in which growth rates are an AR(2) in which two parameters — the intercept and the second AR coefficient — switch between states, and the states display no persistence.

1. Introduction

Applied econometrics is increasingly dominated by non–linear models and estimation techniques. The absence of a body of finite sample theory for non–linear models means that applied research must rely either on asymptotic theory or bootstrapping for inference. The primary asymptotic distributional theory for non–linear models runs roughly as follows. In a sufficiently large sample, the estimator nears the true parameter vector. Via a Taylor's expansion, the parameter estimates are equal to their true value, plus the score evaluated at the true value, divided by the second derivative matrix evaluated at median points[1]. The likelihood surface is assumed to be approximately quadratic in this region, so the second derivatives are approximately constant (that is, not a function of the parameters). Since the score is mean zero, if it has positive variance, we can apply a central limit theorem, and conclude that the estimator has an asymptotic multivariate normal distribution.

There appears to be two key assumptions to this argument. First, the likelihood surface must be locally quadratic. By locally, we must interpret this to mean that the likelihood surface is approximately quadratic over the region in which both the null hypothesis and the global optimum lie (with high probability).[2] In fact, this condition is routinely violated in many applications. For example, if some parameters are not identified under the null hypothesis, then the likelihood function is *flat* (with respect to the unidentified parameters) at the optimum. In other cases, the likelihood surface has more than one local optima, and the null hypothesis may not lie

[1]The rows of the matrix are not necessarily evaluated at the same points.

[2]The requirement that the global optimum lie in the locally quadratic region "with high probability" is somewhat circular, since the argument is designed to provide a *distributional* theory. The conventional proof circumvents this problem by appealing to the consistency of the estimator.

on the same "hill" as the global optimum. In this case, the likelihood surface is far from quadratic in the region between the global optimum and the null hypothesis. The second key assumption is that the score must have a positive variance. This condition is violated when the score is identically zero under the null hypothesis, which occurs when the null hypothesis yields a local maximum, minimum, or inflection point.

Some of these problems have been outlined in the literature before, and separate methods proposed for "handling" the distributional theory in these special cases.

Davies (1977, 1987) analyzes the problem of unidentified nuisance parameters. He suggests viewing the test statistic as a function of the nuisance parameter, in order to apply empirical process theory. Davies bounds the maximum of the empirical process using a crossing–point argument. Hansen (1991) extends the empirical process theory to a wider class of estimation problems and test statistics, but instead of bounding the maximum, provides a direct method to compute critical values, using the empirical covariance function of the empirical process.

Lee and Chesher (1986) study the Lagrange multiplier (LM) test in the case of identically zero scores. They suggest examining higher–order derivatives at the null. This may be useful if the higher–order derivatives are also not identically zero; but even if they are not, the power of their test is not clear.

Each of the above papers present methods which are useful in certain special cases. No general results appear to exist. In an attempt to fill this void, this paper takes a new approach to testing which does not require either that the likelihood be locally quadratic or that the scores (or any other derivative) have positive variance. We work directly with the likelihood surface, viewing the likelihood function as an empirical process of the unknown parameters. Empirical process theory is used to derive a bound for the asymptotic distribution of a standardized likelihood ratio statistic. The distribution depends upon the covariance function of the empirical process associated with the likelihood surface, but we show that the distribution of this empirical process can be easily obtained via simulation.

This testing apparatus is set to work on the Markov trend model of output proposed by Hamilton (1989). Hamilton modeled postwar U.S. GNP growth rates as the sum of an AR(4) process and a Markov process. This may be interpreted as a model where one of the parameters (the mean) switches between two values according to a Markov transition process. Hamilton argued that this model was a better description of the data than the traditional AR model with a fixed mean. As recognized in his original paper, however, this model is plagued by not just one, but *all* of the problems mentioned above. Two nuisance parameters — the transition probabilities — are not identified under the null hypothesis. The null hypothesis also yields a local optimum of the likelihood surface, and higher order derivatives also appear to be zero. This yields a singular information matrix under the null. Being highly non–linear, the model produces numerous local optima as well. Recognizing the inapplicability of standard theory, Hamilton (1989) did not attempt a formal hypothesis test of the null of an AR(4) versus his Markov trend model.

The standardized LR test, which is a valid statistical test to discriminate between these models, fails to reject the null of an AR(4) in favor of the Markov trend model. Apparently, the presence of the two nuisance parameters gives the likelihood surface sufficient freedom so that we cannot reject the possibility that the apparent "significant" coefficients could simply be due to sampling variation.

Hamilton's Markov trend model, however, is quite restrictive in only allowing one parameter to vary with the Markov state. We find strong evidence for an alternative model, in which growth rates are modeled as an AR(2), with the intercept and second AR parameter varying between states. Further, there is no persistence in the states, as the model accepts the restriction that the probability of being in one state or the other is independent of the current state. That is, we find a *mixture* model, rather than a *markov trend* model for GNP. Applying the standardized LR test, this mixture model rejects the null hypothesis of an AR(2) at the 1% level.

Section 2 presents the main theoretical results in a simplified environment without nuisance parameters. Technical details are deemphasized in favor of intuition. Section 3 outlines the theory more completely, allowing for nuisance parameters (both identified and non–identified). These sections develop the apparatus to analyze the likelihood function as an empirical process. Sections 4 and 5 use these methods to analyze post–war quarterly U.S. GNP. Section 4 analyzes the Markov trend model, and section 5 proposes an alternative mixture model. A conclusion follows.

Concerning notation, the symbol "$\Longrightarrow$" is used to denote weak convergence of probability measures with respect to the uniform metric, and "$\|\cdot\|$" is used to denote the Euclidean metric. All limits are taken as the sample size, n, tends to positive infinity.

2. The Likelihood Surface as an Empirical Process

Let us start with a relatively simple problem. Take a log–likelihood function (divided by sample size) which is a function of an unknown parameter α , can be written in the form

$$L_n(\alpha) = \frac{1}{n} \sum_{i=1}^{n} l_i(\alpha) ,$$

with the null and alternative hypotheses

$$H_0: \alpha = \alpha_0 , \qquad H_1 : \alpha \neq \alpha_0 .$$

The likelihood ratio (LR) function is

$$LR_n(\alpha) = L_n(\alpha) - L_n(\alpha_0) = \frac{1}{n} \sum_{i=1}^{n} \left[l_i(\alpha) - l_i(\alpha_0) \right] .$$

The LR surface can be decomposed into its mean, and deviation from mean:

(1) $$LR_n(\alpha) = LR(\alpha) + Q_n(\alpha)$$

where

$$LR(\alpha) = E\left[LR_n(\alpha) \right] ,$$

is the mean, and

$$Q_n(\alpha) = LR_n(\alpha) - LR(\alpha) = \frac{1}{n} \Sigma_1^n q_i(\alpha),$$

is the deviation from the mean, with

$$q_i(\alpha) = \left[l_i(\alpha) - l_i(\alpha_0) \right] - E\left[l_i(\alpha) - l_i(\alpha_0) \right] .$$

It is useful to reflect upon decomposition (1). Under the null hypothesis, the mean $LR(\alpha)$ is non–positive, and strictly negative for $\alpha \neq \alpha_0$. The deviation from the mean, $Q_n(\alpha)$, converges (under mild conditions) pointwise to the 0 function, often uniformly. Note that the unrestricted maximum likelihood estimate of α is also the value of α which maximizes the LR surface $LR_n(\alpha)$. In finite samples, the global optimum will not equal α_0 due to the presence of $Q_n(\alpha)$. Indeed, random fluctuations in the function $Q_n(\alpha)$ are the reason why $LR_n(\alpha)$ is maximized at some value of α other than α_0. We can therefore find some insight into the behavior of the optimization problem by studying the stochastic process $Q_n(\alpha)$. When properly standardized, we find

$$\sqrt{n}\, Q_n(\alpha) \;=\; \frac{1}{\sqrt{n}} \Sigma_1^n\, q_i(\alpha) \;\Rightarrow\; Q(\alpha)\ , \tag{2}$$

where $Q(\alpha)$ is a mean zero Gaussian process with covariance function

$$K(\alpha_1,\alpha_2) \;=\; E\Big[q_i(\alpha_1)q_i(\alpha_2)\Big]\ . \tag{3}$$

The empirical process result (2) is a natural generalization of the classical central limit theorem. For each value of α, $Q(\alpha)$ is a normal random variable with mean zero and variance $K(\alpha,\alpha)$. The function $K(\cdot)$ describes the covariances between $Q(\alpha)$ at different values of α.

The decomposition (1) can be rewritten as an asymptotic approximation:

$$\begin{aligned} \sqrt{n}\, LR_n(\alpha) \;&=\; \sqrt{n}\, LR(\alpha) \;+\; \sqrt{n}\, Q_n(\alpha) \\ &=\; \sqrt{n}\, LR(\alpha) \;+\; Q(\alpha) \;+\; o_p(1) \end{aligned} \tag{4}$$

where the $o_p(1)$ term holds uniformly in α. (4) states that the LR surface equals (in large samples) the mean function plus a Gaussian process. The Gaussian process $Q(\alpha)$ is completely determined by the covariance function $K(\cdot)$ in (3), which can be estimated from the data (we will discuss this in section 3.2). The mean function, $LR(\alpha)$ is unknown. Standard asymptotic theory requires that $LR(\alpha)$ is well–behaved. We can avoid this requirement by instead appealing to the fact that $LR(\alpha) \leq 0$ for all α when the null hypothesis is true. This gives

$$\sqrt{n}\, LR_n(\alpha) \;\leq\; \sqrt{n}\, Q_n(\alpha) \;\Rightarrow\; Q(\alpha)\ . \tag{5}$$

From (5), we can find a bound for the asymptotic distribution of the standard LR test of H_0 against H_1. Since $LR_n = \sup_\alpha nLR_n(\alpha)$, we have as $n \to \infty$,

$$(6) \qquad P\{\tfrac{1}{\sqrt{n}}LR_n \geq x\} \leq P\{\sup_\alpha \sqrt{n}\, Q_n(\alpha) \geq x\} \longrightarrow P\{\sup_\alpha Q(\alpha) \geq x\}.$$

While an interesting theoretical observation, it is not clear that (6) provides a distributional bound which is the most useful in practice. The process $Q(\alpha)$ is Gaussian, but it it not standardized. As $\alpha \to \alpha_0$, for example, $LR_n(\alpha)$ and $Q(\alpha)$ vanish. It seems more sensible to standardize the likelihood ratio so that all values of α yield the same variance. Specifically, noting that the variance function associated with the covariance function (3) is

$$V(\alpha) = K(\alpha,\alpha),$$

with its sample analog,

$$\begin{aligned} V_n(\alpha) &= \frac{1}{n} \Sigma_1^n \left[l_i(\alpha) - l_i(\alpha_0) - LR_n(\alpha)\right]^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[l_i(\alpha) - l_i(\alpha_0)\right]^2 - LR_n(\alpha)^2, \end{aligned}$$

we then have

$$\sqrt{n}\, Q_n(\alpha)/V_n(\alpha)^{1/2} \Longrightarrow \frac{Q(\alpha)}{V(\alpha)^{1/2}} = Q^*(\alpha) \quad \text{, say}.$$

The Gaussian process $Q^*(\alpha)$ has unit variance for all α. Thus for any given α, $Q^*(\alpha) \equiv N(0,1)$. Now define the standardized likelihood ratio process

$$LR_n^*(\alpha) = \frac{LR_n(\alpha)}{V_n(\alpha)^{1/2}}$$

and the standardized likelihood ratio statistic

$$LR_n^* = \sup_\alpha \sqrt{n}\, LR_n^*(\alpha).$$

We now conclude our discussion with the bound

$$P\{LR_n^* \geq x\} \leq P\{\sup_\alpha \frac{\sqrt{n}\ Q_n(\alpha)}{V_n(\alpha)^{1/2}} \geq x\}$$

$$\longrightarrow P\{\sup_\alpha Q^*(\alpha) \leq x\} = F^*(x) .$$

As we show later, we can obtain good approximations to the distribution function $F^*(x)$. One negative feature of this approach is that the asymptotic distribution obtained is a <u>bound</u>. Thus tests based on this approach may be conservative (underrejection when the null is true), and hence suffer a loss in effective power (ability to reject the null when it is false). Given that there appears to be no other general alternative available, this may be a mild criticism at this point.

In some cases, the inequality will be an equality, eliminating this concern. Consider the simple location model : y_t iid $N(\alpha,1)$ with $H_0 : \alpha = 0$ vs. $H_1 : \alpha > 0$. Here

$$LR_n(\alpha) = -\frac{1}{2}\frac{1}{n}\Sigma_1^n\left[(y_t - \alpha)^2 - y_t^2\right] = \alpha\,\bar{y} - \alpha^2/2 .$$

so

$$V_n(\alpha) = \frac{1}{n}\Sigma_1^n\left[\alpha y_t - \alpha^2/2 - (\alpha\bar{y} - \alpha^2/2)\right]^2 = \alpha^2\,\hat{\sigma}_y^2$$

and thus

$$LR^*{}_n(\alpha) = \sqrt{n}\ (\bar{y} - \alpha/2)/\hat{\sigma}_y \quad .$$

We find that

$$LR^*{}_n = \sup_\alpha \sqrt{n}\ LR_n^*(\alpha) = \sqrt{n}\ \bar{y}/\hat{\sigma}_y$$

which is the standard t–statistic for the test of H_0 against H_1 . In this simple example, the standardized LR statistic has a conventional interpretation and distribution. This will not always be the case, but it suggests that the structure of the standardized LR statistic is not as unconventional as appears at first glance.

3. General Theory

3.1 Allowing for Nuisance Parameters

The previous section was meant to be motivational, since most problems of interest contain nuisance parameters. Suppose that the model has log–likelihood

$$L_n(\beta,\gamma,\theta) = \frac{1}{n} \Sigma_1^n l_i(\beta,\gamma,\theta)$$

with parameter vectors $\beta \in B$, $\gamma \in \Gamma$, and $\theta \in \Theta$. The hypothesis takes the form

$$H_0 : \beta = 0 \qquad\qquad H_1 : \beta \neq 0 .$$

Note that θ and γ are nuisance parameters. Assume that θ is fully identified, but γ is not identified under H_0 . (This requires that $L_n(0,\gamma,\theta)$ not depend upon γ). In order to apply a testing method similar to that suggested in section 2, we are going to have to eliminate the parameter vector θ. We do this by concentration.

Set $\alpha = (\beta' , \gamma')'$ and $A = B \times \Gamma$, and $L_n(\alpha,\theta)$ and $l_i(\alpha,\theta)$ accordingly. Define the sequence of parameter estimates

$$\hat{\theta}(\alpha) = \underset{\theta \in \Theta}{\text{Max}}\ L_n(\alpha,\theta) \tag{7}$$

which are the maximum likelihood estimates of θ for fixed values of α and γ . The concentrated likelihood function is then

$$\hat{L}_n(\alpha) = L_n(\alpha,\hat{\theta}(\alpha)) .$$

Ideally, we would like to be working with the large–sample concentrated likelihood function given by

$$L_n(\alpha) = L_n(\alpha,\theta(\alpha))$$

where

$$\theta(\alpha) \;=\; \underset{\theta\in\Theta}{\operatorname{Argmax}} \lim_{n\to\infty} \; E\, L_n(\alpha,\theta)$$

is the pseudo–true value of θ, for fixed α. In order for the concentration argument to work, we require that $\hat{\theta}(\alpha)$ is consistent for $\theta(\alpha)$ at rate $\sqrt{n}$, uniformly in α. Set $D(\alpha) = \hat{\theta}(\alpha) - \theta(\alpha)$. Formally, we assume

$$\text{(A1)} \qquad \sup_{\alpha\in A} \sqrt{n}\, \| D(\alpha) \| \;=\; O_p(1) .$$

In order to show (A1) from more primitive assumptions, we would have to assume that the maximization problem given in (7) satisfies the standard assumptions for non–linear estimators. That is, we are assuming that all of the "trouble" arises in the parameters $\alpha = (\beta',\gamma')$. We further require that the matrix of second derivatives with respect to θ be well behaved. If we define

$$M_n(\alpha,\theta) \;=\; \frac{\partial^2}{\partial\theta\partial\theta'} L_n(\alpha,\theta) ,$$

we require

$$\text{(A2)} \qquad \sup_{\alpha\in A,\,\theta\in\Theta} \| M_n(\alpha,\theta) \| \;=\; O_p(1) .$$

By a Taylor's expansion, we have

$$L_n(\alpha,\theta(\alpha)) - L_n(\alpha,\hat{\theta}(\alpha)) \;=\; D(\alpha)'\frac{\partial}{\partial\theta} L_n(\alpha,\hat{\theta}(\alpha)) \;+\; \frac{1}{2}\, D(\alpha)' M_n(\alpha,\theta^*(\alpha))\, D(\alpha) ,$$

where $\theta^*(\alpha)$ lies on a line segment joining $\hat{\theta}(\alpha)$ and $\theta(\alpha)$. This gives

$$\text{(8)} \qquad \sup_{\alpha\in A} \|L_n(\alpha) - \hat{L}_n(\alpha)\| \;=\; \sup_{\alpha\in A} \| D(\alpha)' M_n(\alpha,\theta^*(\alpha))\, D(\alpha) \| \;=\; O_p(\tfrac{1}{n}) .$$

We now proceed as in section 2. The likelihood ratio process, its large–sample counterpart, expectation, and centered versions are

$$\hat{L}R_n(\alpha) \;=\; \hat{L}_n(\alpha) \;-\; \hat{L}_n(0,\gamma) ,$$

$$\begin{aligned} LR_n(\alpha) &= L_n(\alpha) - L_n(0,\gamma) , \\ LR(\alpha) &= E\left[LR_n(\alpha)\right] \\ \hat{Q}_n(\alpha) &= \hat{L}R_n(\alpha) - LR(\alpha) \\ Q_n(\alpha &= LR_n(\alpha) - LR(\alpha) . \end{aligned}$$

We now assume that an empirical process central limit theorem (CLT) holds:

(A3) $$\sqrt{n}\ Q_n(\alpha) \Longrightarrow Q(\alpha)$$

where $Q(\alpha)$ is a Gaussian process with covariance function

$$K(\alpha_1;\alpha_2) = \lim_{n\to\infty} n\ E\left[Q_n(\alpha_1)\ Q_n(\alpha_2)\right] .$$

Andrews (1990) recently has provided an empirical process CLT under conditions which permit temporal dependence and heterogeneity. Essentially, the likelihood components $l_i(\alpha,\theta(\alpha))$ need to have bounded $2+\delta$ moments, satisfy a mixing or near epoch dependence condition, and satisfy a smoothness condition with respect to α .

Using the fact that $LR(\alpha) = LR(\beta,\gamma) \leq 0$ under the null hypothesis, (8), and (A3), we obtain a limit theory for the concentrated likelihood process:

$$\sqrt{n}\ \hat{L}R_n(\alpha) \leq \sqrt{n}\ \hat{Q}_n(\alpha) = \sqrt{n}\ Q_n(\alpha) + o_p(1) \Longrightarrow Q(\alpha) .$$

Note that the $o_p(1)$ term holds uniformly in α .

As discussed in the previous section, it makes sense to standardize the LR process. Construct the sample variance

$$V_n(\alpha,\hat{\theta}(\alpha)) = \frac{1}{n}\ \Sigma_1^n q_i(\alpha,\hat{\theta}(\alpha))^2 ,$$

where

$$q_i(\alpha,\hat{\theta}(\alpha)) = l_i(\alpha,\hat{\theta}(\alpha)) - l_i(0,\gamma,\hat{\theta}(0,\gamma)) - \hat{L}R_n(\alpha) .$$

We assume that this estimator is uniformly consistent for $V(\alpha) = K(\alpha;\alpha)$:

(A4) $$\sup_{\alpha\in A} \| V_n(\alpha,\theta(\alpha)) - V(\alpha) \| \to_p 0 .$$

Finally, we need to be able to find some region $A^* \subset A$ such that

(A5) $$\inf_{A\in A^*} V(\alpha) > 0 .$$

From this, we can calculate the standardized LR function and LR statistic

$$\hat{LR}_n^*(\alpha) = \frac{\hat{LR}_n(\alpha)}{V_n(\alpha)^{1/2}} , \qquad \hat{LR}_n^* = \sup_{\alpha\in A^*} \sqrt{n}\, \hat{LR}_n^*(\alpha) ,$$

and the centered stochastic processes

$$\hat{Q}_n^*(\alpha) = \frac{\hat{Q}_n(\alpha)}{V_n(\alpha)^{1/2}} , \qquad Q_n^*(\alpha) = \frac{Q_n(\alpha)}{V_n(\alpha)^{1/2}} .$$

We have

$$\begin{aligned} \hat{LR}_n^* &\le \sup_{\alpha\in A^*} \sqrt{n}\, \hat{Q}_n^*(\alpha) \\ &= \sup_{\alpha\in A^*} \sqrt{n}\, Q_n^*(\alpha) + o_p(1) \\ &\Rightarrow \sup_{\alpha\in A^*} Q^*(\alpha) \equiv \mathrm{Sup}Q^* , \end{aligned}$$

where $Q^*(\alpha)$ is a Gaussian process with covariance function

$$K^*(\alpha_1;\alpha_2) = \frac{K(\alpha_1;\alpha_2)}{V(\alpha_1)^{1/2} V(\alpha_2)^{1/2}}$$

We have shown the following result.

<u>Theorem 1</u>. *Under* (A1) – (A5),

$$P\{\hat{LR}_n^* \ge x\} \le P\{\sup_{\alpha\in A^*} \sqrt{n}\, \hat{Q}_n^*(\alpha) \ge x\} \to P\{\mathrm{Sup}Q^* \ge x\}.$$

Theorem 1 provides a bound for the standardized LR statistic in terms of the distribution of the random variable SupQ* . The assumptions (A1) through (A5) are high–level, but quite weak, in contrast to conventional distributional theory. Thus Theorem 1 is applicable in a much wider class of models than the standard theory. The cost is the presence of the inequality. The fact that the distribution of the test statistic is only bounded means that the test may be conservative and effective power may be lowered. Hence, Theorem 1 should only be used (vis–a–vis conventional theory) when it is apparent that the conventional assumptions are invalid.

3.2 Calculating the Asymptotic Distribution

The distribution of the random variable SupQ* presented in Theorem 1 is generally non–standard, precluding generic tabulation. Following Hansen (1991), it is quite easy, however, to use the empirical covariance function to generate the asymptotic distribution via simulation.

The random variable SupQ* is the supremum of the empirical process $Q^*(\alpha)$, which is completely characterized by its covariance function $K^*(\cdot)$. We do not know K^* , but we have the sample analog

$$\hat{K}_n^*(\alpha_1;\alpha_2) = \frac{\frac{1}{n}\Sigma_1^n \, q_i(\alpha_1,\hat{\theta}(\alpha_1))\, q_i(\alpha_2,\hat{\theta}(\alpha_2))}{V_n(\alpha_1)^{1/2} V(\alpha_2)^{1/2}} \; .$$

Suppose that we can draw iid Gaussian processes whose covariance function is $\hat{K}_n^*(\cdot)$. The supremum of each of these processes (approximately) has the distribution SupQ*, where the approximation is only due to the sample discrepancy between $\hat{K}_n^*$ and K^*, which vanishes in large samples. Through repeated draws from this urn, we can (approximately) obtain the distribution SupQ* from the empirical distribution of the random draws. For example, critical values and p–values can be calculated, and histograms plotted, for any given example.

An easy method to obtain draws from the required family of Gaussian processes is to generate a random sample of N(0,1) variables $\{u_i\}_1^n$, and then construct

$$\tilde{L}R^*(\alpha) \;=\; \frac{1}{\sqrt{n}}\,\frac{\Sigma_1^n q_i(\,\alpha,\hat{\theta}(\alpha))u_i}{V_n(\alpha)^{1/2}}\;.$$

It is straightforward to verify that (conditional on the data) the process $\tilde{L}R^*(\cdot)$ is Gaussian with covariance function $\hat{K}_n^*$. It is also (conditionally) independent from other processes $\tilde{L}R^*(\cdot)$ constructed with independent samples $\{u_i\}$. It is evident that this construction meets our requirements.

3.3 Practical Issues

The main cost of the procedures advocated here is in the evaluation of the likelihood across different values of $\alpha = (\beta,\gamma)$. The need to concentrate out the identified nuisance parameters (θ) means that for each value of α , the constrained likelihood needs to be optimized. This can be a major computational burden, even if the parameter space is small.

As far as I can see, the only practical way to evaluate the maximal statistics discussed here is to form a grid search over a relatively small number of values of α . A trade–off arises as a more extensive grid search requires more computation, but reduces the arbitrariness associated with the choice of grid, and may increase the power of the test. For every value of α at which the constrained likelihood is optimized, one needs to calculate only the sequence $\{q_i(\alpha,\hat{\theta}(\alpha))\}$ (an $n\times 1$ vector). Storage requirements are therefore equal to the number of grid points multiplied by sample size. From these numbers, both the modified LR statistic and its asymptotic distribution can be calculated.

4. Testing the Markov Trend Model of GNP

What is a good univariate model of GNP? Since the degree of persistence in linearly detrended GNP is quite high (seemingly non–stationary), but the amount of persistence in growth rates is relatively low, we will be interested in finding a model for the first difference of the natural log of real GNP, which we will denote by x_t .

A reasonable starting place is the autoregressive (AR) model:

$$\varphi(L)x_t = \mu + e_t , \tag{9}$$

where e_t is iid , perhaps from a normal distribution. The argument for the AR model is that (practically) all covariance stationary processes have an autoregressive representation, which can be written as (9) where the error e_t is white noise. The reasonableness of adopting the AR model is that most of the estimation and inference techniques designed for the AR model are valid under the broader conditions of an AR representation, so an applied researcher need not be worried that they have the "wrong" model. The results of fitting an AR(4) to post–war quarterly U.S. GNP are presented in Table 1.[3]

The representation argument does not imply that an AR model is adequate for all purposes. A Gaussian AR model is incompatible, for example, with the observed asymmetry between expansions and contractions. This asymmetry could be "explained" by an AR model with skewed innovations e_t , but this solution is not completely satisfactory. If, for instance, the errors in the AR representation are not independent, but have conditionally forecastable third moments, then the AR model is suboptimal, since it is not taking into account forecastable asymmetries in the business cycle.

Many alternatives to the AR model are possible. Hamilton (1989) proposed a

[3]All regressions are reported with heteroskedasticity–consistent standard errors (see White, 1980). Also reported is the Gaussian log–likelihood and the value of the LM test for parameter instability proposed in Nyblom (1989) and Hansen (1990).

"Markov Trend" formulation in which a large degree of explanatory power is assigned to the existence of a few "states" between which the economy shifts according to a Markov process. For GNP growth rates, Hamilton suggested the model

$$x_t = \mu + \mu_d s_t + u_t , \qquad \varphi(L)u_t = e_t ,$$

where s_t is a dummy variable equaling either 1 or 0 . The transitions between these states are governed by the transition probabilities

$$P\{s_t = 1 \mid s_{t-1} = 1\} = p$$

$$P\{s_t = 0 \mid s_{t-1} = 0\} = q .$$

In his paper, Hamilton set the autoregressive order equal to four. In order to estimate the model by maximum likelihood, Hamilton added the assumption that e_t is iid $N(0,\sigma^2)$ and independent of $\{s_t\}$.

Table 2 reports estimates for this Markov trend model. The estimates look reasonable and significant. Notice that the heteroskedasticity–consistent standard error estimates are larger than the conventional standard error estimates reported in Hamilton (1989).

The Markov trend model reduces to the AR(4) under the constraint

$$H_0 : \mu_d = 0 .$$

To test this hypothesis, one would be tempted to either compute the likelihood ratio statistic from Tables 1 and 2, or the t–statistic for μ_d from Table 2. These test statistics, however, do not have a standard distributional theory. Two reasons are paramount. First, under the null hypothesis, the transition probabilities p and q are not identified. As mentioned in the introduction, this means that the large sample likelihood surface is flat (under the null) with respect to these parameters. The asymptotic likelihood has no unique maximum and is not locally quadratic. Second, the scores with respect to μ_d , p , and q are identically zero when evaluated at the null hypothesis. A bit of experimentation indicated that higher order derivatives were also zero. The combination of these problems suggests that standard distributional theory is inapplicable. As an alternative, I will use the generalized testing procedure presented in section 2.

In the notation of section 2, $\beta = \mu_d$, $\gamma = (p, q)$, and $\theta = (\mu , \sigma^2, \varphi_1, \varphi_2, \varphi_3, \varphi_4)$. The test requires computing the constrained estimates of θ for each combination of $\alpha = (\mu_d , p , q)$ for some grid of values. For μ_d , I used the range [.1 , 2] in steps of .1 ; and for p and q , the range [.15, .90] in steps of .15 . This requires estimation of the concentrated likelihood at 720 points. That is, the six parameters in θ must be found numerically for each value of μ_d , p and q . In order to achieve some efficiency in this estimation, for each value of p and q , I started with $\mu_d = 0.1$, and used for starting values the null estimates (which correspond to $\mu_d = 0$). After convergence was obtained, I moved on to $\mu_d = .2$, and used for starting values the final values from the previous optimization, and so on. This kept the computation time down to a reasonable degree, and seemed to produce the correct results. For example, if p and q are fixed at the global optimum (from Table 3), and you do this procedure, you obtain the correct value of the likelihood when μ_d achieves its global estimate.

The value of the standardized LR statistic was calculated as outlined in section 3.1, and found to be 1.47.[4] If a standard normal theory were applicable, this statistic would not reject the null hypothesis at the 5% level based on the one–sided critical values, but it would be close. The standard normal theory, however, is <u>not</u> applicable, since the standardized likelihood surface has been maximized over 720 points!

To calculate the asymptotic distribution, we can use the method of section 3.2.

---

[4]The standardized LR function was maximized at $\mu_d = 1.1$, p = .90, and q = .60.

Computationally, this takes much less time than obtaining the constrained estimates. 1000 random samples were drawn and used to construct the bounding random variable. The value 1.47 is far from significant, with a p–value of 0.82 . The approximate 5% critical value is 3.1. The result is unambiguous. The AR(4) model is not rejected, and the statistical technique fails to find any evidence in favor of the Markov trend model.

The density of the bounding asymptotic distribution was estimated by a normal kernel and is displayed in Figure 1. Not surprisingly, the density is significantly different from a standard normal, and is skewed to the right.

There are two different sources of this non–normality. One is the presence of nuisance parameters, and the other is the maximization over the structural parameter μ_d . In order to get a sense of their relative contributions to the distributional problem, I tried to separate out their effects by doing the same calculations, while holding fixed a subset of the parameters. First, I fixed the transition probabilities at the values which maximize the standardized LR function, and calculated the asymptotic distribution of the test statistic as if these were known *a priori*. This should give an approximate feel for the contribution of maximizing over μ_d alone. By this calculation, the 5% critical value drops from 3.10 to 2.2 and the observed value of 1.47 yields a p–value of 0.21. Second, I fixed μ_d at the value which maximized the standardized LR function, (pretending as if this value were *a priori* known) and calculated the asymptotic distribution via maximization over p and q . This yields a 5% critical value of 2.63, and a p–value for 1.47 of 0.52. These calculations are not rigorous, but suggest that the thick tail of the density shown in figure 1 is primarily the result of the two unidentified nuisance parameters, rather than the maximization over α_d. (Note for contrast that if all three parameters are fixed, then the calculation yields an exact normal distribution, where the 5% critical value is 1.65.)

5. A Mixture AR Model of GNP

In Hamilton's Markov trend model of GNP, the difference between states of the world is completely captured by differences in the mean of the process. In the modified Markov trend model discussed at length in section 4, the difference between states is contained in differences in the intercept. It seems odd to impose this restriction *a priori*. Delong and Summers (1988), for example, argue that during the Great Depression, shocks to GNP were more persistent. They suggest that shifting autoregressive parameters can capture this phenomenon.

Letting the autoregressive parameters shift between states would be computationally difficult in the model Hamilton estimates. A computationally simple alternative is to modify Hamilton's model from an AR(4) with a switching mean to an AR(4) with a switching intercept. This modified model is

$$\varphi(L)x_t = \mu + \mu_d s_t + e_t ,$$

where s_t is defined as before. In the results for the Markov trend model (Table 2) most of the variation is picked up by the Markov process, and relatively little is accounted for by the autoregression. Thus we would expect these two models to perform similarly in the present context.[5]

The estimates from this modified Markov trend model are given in Table 3. The point estimates are quite similar to those in Table 2, but the standard errors are generally smaller and the log–likelihood is higher. It appears that the modified model performs even better than Hamilton's model, although the difference is probably not statistically significant[6]. The standardized LR statistic is 1.95 for this model, with a

[5]As a general rule, the two models have quite different dynamics, as pointed out by Hamilton (1991).

[6]One could use the methods of Vuong (1989) to report a formal comparison of the models, but it doesn't appear to be worth the effort.

p–value of .44, which is still quite far from significant.

We can now relax the assumption that only the intercept varies between states. The first two columns in Table 4 (under "Unconstrained") report the estimates from estimation of a fully unrestricted model, in which the intercept, slope parameters, and error variance are all allowed to shift between the two states. All of the parameters with the "d" subscript denote the *difference* in coefficients between states.

These estimates give a very different picture from the Markov model of table 3. The restrictions implied by the shifting intercept model are rejected by a Wald test at the 1% level. There appears to be a second shifting parameters, the second AR lag. While the Markov trend model of table 3 yielded small point estimates for the first two AR parameters, the unconstrained estimates are much larger, while the third and fourth AR parameters are quite small.

The transition probabilities (p and q) also tell a different story. The unconstrained estimates are much smaller, and sum to 1.026. This suggests that the constraint $q = 1-p$ should be satisfied. This constraint is of importance for two reasons. First, it eliminates one unidentified nuisance parameter, making the testing problem better behaved. Second, the model has a different interpretation. $p + q = 1$ implies that there is *no persistence* in the Markov process, for the probability that s_t takes on one or zero is independent of the previous state. It seems appropriate to call this model a *Mixture Model* of GNP, rather than a Markov trend model, since the parameters are varying according to a mixture distribution, with no persistence.

Testing all of these seven restrictions yields a Wald statistic of only 1.07. Estimates of the restricted model are reported in the last two columns of Table 4 (under "Mixture Model"). As expected, the parameter estimates are very close to the unconstrained estimates, but the estimated standard errors are much smaller.

The elimination of one unidentified nuisance parameter and the great increase in fit suggest that testing this mixture model against an autoregressive model might succeed. Since the third and fourth autoregressive parameters have been eliminated from the mixture model (and are not significant in the AR(4)), it appears to make sense to test the mixture model against an AR(2), increasing the sample size by two. We now have two structural parameters, μ_d and φ_{2d}, and one unidentified nuisance parameter, p. The identified nuisance parameters are μ, σ, φ_1, φ_2 (two less than before). To calculate the concentrated likelihood function, I used the following grid. μ_d from .2 to 2, in steps of .2; φ_{2d} from -1 to 1 in steps of .2; and p from .15 to .90 in steps of .15.

The standardized LR test statistic is 4.93. The asymptotic distribution, calculated with 1000 replications of normal random samples, yields an upper 5% critical value of 3.09, and the test statistic is found to significant at the asymptotic 1% level. Thus we are able to reject the AR(2) in favor of the mixture model of GNP at a high level of statistical significance.

6. Conclusion

This paper has set out to develop a method of hypothesis testing for non–linear models which do not necessarily satisfy the standard list of regularity conditions. With the growing popularity of non–linear models, more attention should be paid to regularity conditions and their violation. Statistical tools to conduct inference when regularity conditions are violated are noticeably absent.[7] This paper proposes a new and quite different approach to the subject. Essentially, the suggestion is to view the likelihood surface as the sum of the limit function and an empirical process. Random variation in estimation is entirely due to the interplay between the limit function and the random empirical process. While all we may know about the limit of the likelihood surface is that it is maximized at the null value, we can calculate the asymptotic distribution of the likelihood empirical process from the data itself. This enables us to bound the distribution of the maximum of the standardized likelihood ratio process, and use this maximum as test of the null hypothesis.

This paper also investigates the statistical significance of Hamilton's (1989) Markov trend model of GNP. The violations of the conventional regularity conditions are strong, and I am unable to reject the hypothesis that the "good fit" of Hamilton's model is simply due to sampling error. Instead, I estimate an alternative which I call a mixture model of GNP, which is an AR(2) where the intercept and the second AR parameter randomly shift between two values. This mixture model fits the data better than an AR(2), rejecting the latter at the asymptotic 1% level.

[7]The one issue which has been discussed at length is estimation and testing subject to boundary conditions. See, for example, Chernoff (1954), Moran (1971), Gourieroux, et. al. (1982), Rogers (1986), and Wolak (1989).

Table 1

Maximum Likelihood Estimates of Gaussian AR Model
Based on Data for U.S. Real GNP, t = 1952:2 to 1984:4

Parameter	Estimate	Standard Error
μ	0.557	0.140
φ_1	0.310	0.085
φ_2	0.127	0.095
φ_3	–0.121	0.087
φ_4	–0.089	0.090
σ	0.983	0.064

Log–Likelihood:	–183.669	
LM Stability Test:	0.958	(Insignificant at 20% level)

Table 2

Maximum Likelihood Estimates of Hamilton Markov Trend Model
Based on Data for U.S. Real GNP, t = 1952:2 to 1984:4

Parameter	Estimate	Standard Error
μ	–0.359	0.465
μ_d	1.522	0.464
φ_1	0.013	0.164
φ_2	–0.058	0.219
φ_3	–0.247	0.148
φ_4	–0.213	0.136
σ	0.769	0.094
p	0.904	0.033
q	0.755	0.101

Log–Likelihood	–181.263	
LM Stability Test	1.364	(Insignificant at 20% level)
Standardized LR Test	1.47	
(p–value)	0.82	

Table 3

Maximum Likelihood Estimates of Modified Markov Trend Model
Based on Data for U.S. Real GNP, t = 1952:2 to 1984:4

Parameter	Estimate	Standard Error
μ	−0.447	0.305
μ_d	1.560	0.245
φ_1	0.112	0.105
φ_2	0.065	0.081
φ_3	−0.126	0.080
φ_4	−0.136	0.091
σ	0.789	0.066
p	0.912	0.032
q	0.669	0.143

Log–Likelihood	−180.184	
Stability Test	0.954	(Insignificant at 20% level)
Standardized LR Test	1.95	
(p–value)	0.44	

Table 4

Maximum Likelihood Estimates of Modified Markov Trend Model
Based on Data for U.S. Real GNP, t = 1952:2 to 1984:4

	Unconstrained		Mixture Model	
Parameter	Estimate	Std. Error	Estimate	Std. Error
μ	–0.690	0.400	–0.756	0.169
μ_d	1.815	0.362	1.871	0.158
φ_1	0.321	0.211	0.321	0.079
φ_2	0.510	0.228	0.461	0.115
φ_3	–0.078	0.121		
φ_4	–0.022	0.148		
φ_{1d}	–0.005	0.215		
φ_{2d}	–0.596	0.153	–0.582	0.133
φ_{3d}	0.006	0.189		
φ_{4d}	0.010	0.356		
σ	0.657	0.121	0.650	0.078
σ_d	0.013	0.255		
p	0.638	0.471	0.619	0.072
q	0.388	0.299		
Likelihood		–174.388		–176.990
Stability		1.463		0.624
Standardized LR Test				4.93
(p–value)				0.00

Table 5 : Wald Tests

Test	Statistic	D.O.F.	p–value
Unconstrained vs. Modified Markov Trend	15.8	5	0.007
Unconstrained vs. Mixture Model	1.1	7	0.994

References

Andrews D.W.K. (1990): "An empirical process central limit theorem for dependent non–identically distributed random variables," Yale University.

Chernoff, H. (1954): "On the distribution of the likelihood ratio," *Annals of Mathematical Statistics*, 25, 573–578.

Davies, R.B. (1977): "Hypothesis testing when a nuisance parameter in present only under the alternative," *Biometrika*, 64, 247–254.

Davies, R.B. (1987): "Hypothesis testing when a nuisance parameter in present only under the alternative," *Biometrika*, 74, 33–43.

DeLong, J.B. and L.H. Summers (1988): "On the existence and interpretation of a 'unit root' in U.S. GNP," *NBER Working Paper Series*, No. 2716.

Gourieroux, C., A. Holly and A. Monfort (1982): "Likelihood ratio test, Wald test and Kuhn–Tucker test in linear models with inequality constraints on the regression parameters, *Econometrica*, 50, 63–80.

Hamilton, J.D. (1989): "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, 57, 357–384.

Hamilton, J.D. (1991): "Estimation, inference, and forecasting of time series subject to changes in regime," *Handbook of Econometrics*, vol. 10, (forthcoming).

Hansen, B.E. (1990): "Lagrange multiplier tests for parameter instability in non–linear models," University of Rochester.

Hansen, B.E. (1991): "Inference when a nuisance parameter is not identified under the null hypothesis," University of Rochester.

Lee, L.F. and A. Chesher (1986): "Specification testing when score test statistics are identically zero," *Journal of Econometrics*, 31, 121–149.

Moran, P.A.P. (1971): "Maximum–likelihood estimation in non–standard conditions," *Proceedings of the Cambridge Philosophical Society*, 70, 441–450.

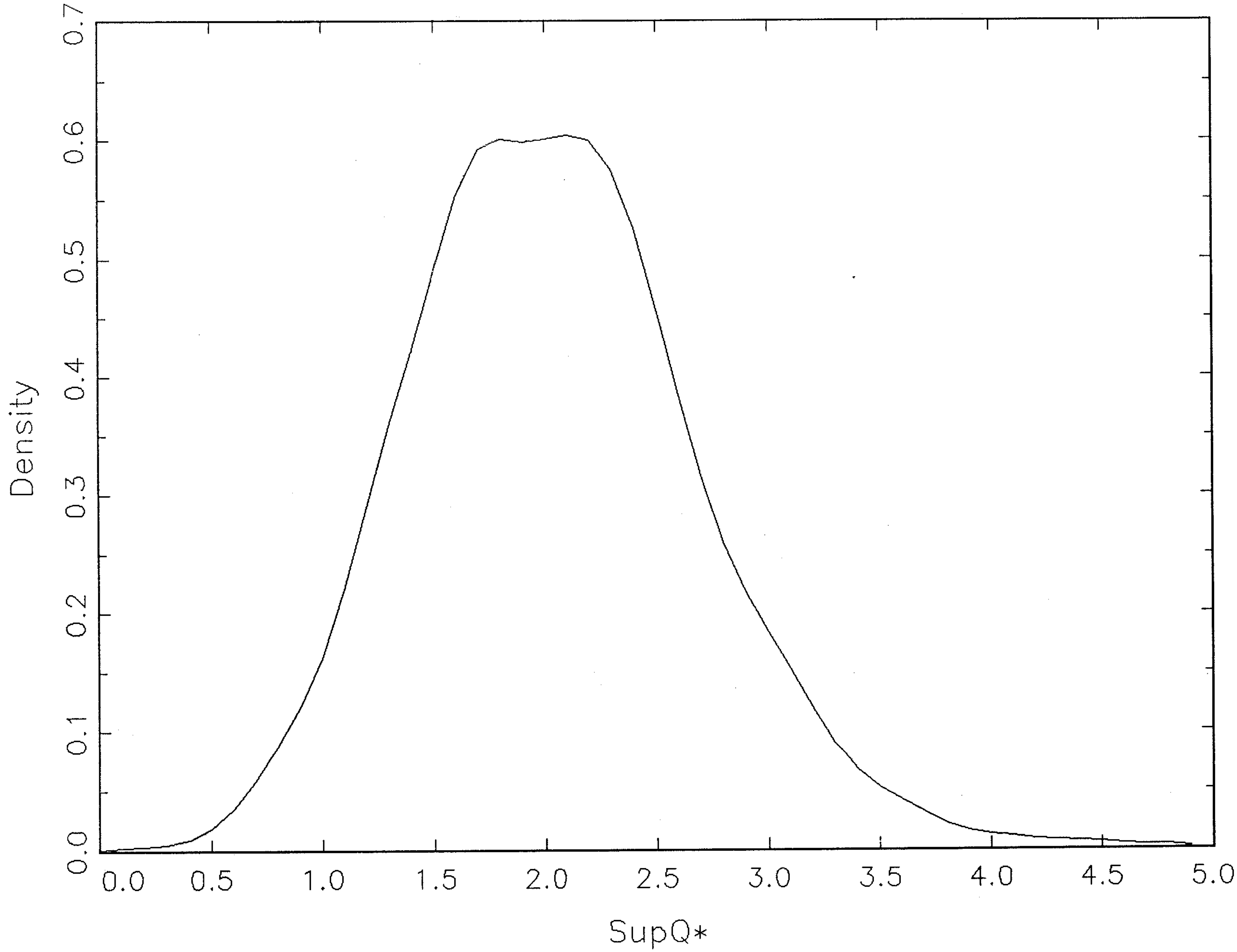
Nyblom, J. (1989): "Testing for the constancy of parameters over time," *Journal of the American Statistical Association*, 84, 223–230.

Rogers, A.J. (1986): "Modified Lagrange multiplier tests for problems with one–sided alternatives," *Journal of Econometrics*, 31, 341–361.

Vuong, Q.H. (1989): "Likelihood ratio tests for model selection and non–nested hypotheses," *Econometrica*, 57, 307–333.

White, H. (1980): "A heteroskedastic–consistent covariance matrix estimator and a direct test for heteroskedasticity", *Econometrica*, 48, 817–838.

Wolak, F.A. (1989): "Local and global testing of linear and nonlinear inequality constraints in nonlinear econometric models," *Econometric Theory*, 5, 1–35.



Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SupQ*