Inferring the Predictability Induced by a Persistent Regressor in a Predictive Threshold Model

Jesùs Gonzalo
Universidad Carlos III de Madrid
jesus.gonzalo@uc3m.es

Jean-Yves Pitarakis
University of Southampton
j.pitarakis@soton.ac.uk

April, 2015

Abstract

We develop tests for detecting possibly episodic predictability induced by a persistent predictor. Our framework is that of a predictive regression model with threshold effects and our goal is to develop operational and easily implementable inferences when one does not wish to impose à priori restrictions on the parameters of the model other than the slopes corresponding to the persistent predictor. Differently put our tests for the null hypothesis of no predictability against threshold predictability remain valid without the need to know whether the remaining parameters of the model are characterised by threshold effects or not (e.g. shifting versus non-shifting intercepts). One interesting feature of our setting is that our test statistics remain unaffected by whether some nuisance parameters are identified or not. We subsequently apply our methodology to the predictability of aggregate stock returns with valuation ratios and document a robust countercyclicality in the ability of some valuation ratios to predict returns in addition to highlighting a strong sensitivity of predictability based results to the time period under consideration.

Keywords: Predictive Regressions, Threshold Effects, Predictability of Stock Returns.

1 Introduction

Predictive regressions are simple regression models in which a highly persistent variable is used as a predictor of a noisier time series. The econometric difficulties that arise due to the combination of a persistent regressor and possible endogeneity have generated an enormous literature aiming to improve inferences in such settings. Common examples include the predictability of stock returns with valuation ratios, the predictability of GDP growth with interest rates amongst numerous others (see for instance Valkanov (2003), Lewellen (2004), Campbell and Yogo (2006), Jansson and Moreira (2006), Rossi (2007), Bandi and Perron (2008), Ang and Bekaert (2008), Wei and Wright (2013) and more recently Kostakis, Magdalinos and Stamatogiannis (2014)).

In a recent paper Gonzalo and Pitarakis (2012) have extended the linear predictive regression model into one that allows the strength of predictability to vary across economic episodes such as expansions and recessions. This was achieved through the inclusion of threshold effects which allowed the parameters of the model to switch across regimes driven by an external variable. Within this piecewise linear setting the authors developed a series of tests designed to detect the presence of threshold effects in *all* the parameters of the model by maintaining full linearity within the null hypotheses (i.e. restricting both intercepts and slopes to be stable throughout the sample). Differently put this earlier work was geared towards uncovering regimes within a predictive regression setting rather than determining the predictability of a particular predictor per se.

The goal of this paper is to develop a toolkit that will allow practitioners to test specifically the null hypothesis of no predictability induced by a persistent regressor without restricting the remaining parameters of the model (e.g. intercepts may or may not exhibit threshold effects). Indeed, a researcher may wish to assess the presence of predictability induced solely by some predictor x_t while remaining agnostic about the presence or absence of regimes in the remaining parameters. Moreover, in applications involving return predictability with valuation ratios such as the Dividend Yield and a threshold variable proxying the business cycle, rejection of the null of no predictability on the basis of a null hypothesis that restricts all the parameters of the model as in Gonzalo and Pitarakis (2012) may in fact be driven by the state of the business cycle rather than the predictability induced by the Dividend Yield itself.

The type of inference we consider in this paper naturally raises important identification issues which we address by exploring the feasibility of conducting inferences on the relevant slope parameters that are possibly immune to any knowledge about the behaviour of the intercepts and in particular to whether the latter are subject to regime shifts or not. Our null hypothesis of interest here allows for the possibility of having *nuisance* parameters that may or may not switch across regimes.

Our proposed inferences are based on a standard Wald type test statistic whose distribution we derive under the null hypothesis of no predictability induced by a highly persistent regressor. The limiting distribution of our test statistic evaluated at a particular location of the threshold parameter is then shown to be immune to whether the remaining parameters of the model shift or not. Since the limiting distribution in question depends on a series of nuisance parameters it is not directly usable for inferences unless one wishes to impose an exogeneity assumption on the predictor. Using an Instrumental Variable approach we propose a modified Wald statistic whose new distribution is shown to be standard and free of nuisance parameters under a very general setting.

The plan of the paper is as follows. Section 2 presents our operating model and the underlying probabilistic assumptions. Section 3 develops the large sample inferences. Section 4 illustrates their properties and usefulness via a rich set of simulations. Section 5 applies our proposed methods to the predictability of aggregate US equity returns using a wide range of valuation ratios and Section 6 concludes.

2 The Model and Assumptions

We operate within the same setting as in Gonzalo and Pitarakis (2012). Our predictive regression model with threshold effects or *Predictive Threshold Regression* (PTR) is given by

$$y_{t+1} = (\alpha_1 + \beta_1 x_t) I(q_t \leq \gamma) + (\alpha_2 + \beta_2 x_t) I(q_t > \gamma) + u_{t+1} \tag{1}$$

where the highly persistent predictor x_t is modelled as the nearly integrated process

$$x_t = \rho_T x_{t-1} + v_t, \qquad \rho_T = 1 - \frac{c}{T} \tag{2}$$

with $c > 0$ and $q_t = \mu_q + u_{qt}$ denoting the stationary threshold variable. It is useful to reformulate (1) in matrix form as $y = Z(\gamma)\ \theta + u$ with $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2)$, $Z(\gamma) = (X_1(\gamma)\ X_2(\gamma))$ and with the $X_i(\gamma)$ matrices stacking the elements of $(I(q_t \leq \gamma)\ x_t I(q_t \leq \gamma))$ and $(I(q_t > \gamma)\ x_t I(q_t > \gamma))$ for $i = 1, 2$. We will also use the notation I_{1t} and I_{2t} to refer to $I(q_t \leq \gamma)$ and $I(q_t > \gamma)$ and we let the column vectors x_i and I_i stack the elements $x_t I_{it}$ and I_{it} so that $Z(\gamma) = (I_1\ \ x_1\ \ I_2\ \ x_2)$. Finally and throughout the paper we make use of $I(q_t \leq \gamma) \equiv I(F(q_t) \leq \lambda)$ with $F(.)$ denoting the distribution function of q_t. This allows us to refer to the threshold parameter of interest as γ or λ. Given the assumptions that will be imposed on q_t (e.g. strict stationarity and ergodicity) it is also useful to note that $E[I_{1t}] = \lambda$ and $E[I_{2t}] = 1 - \lambda\ \forall t$. In what follows it will be understood that $\lambda \in \Lambda = [\underline{\lambda}, \overline{\lambda}]$ with $0 < \underline{\lambda} < \lambda < \overline{\lambda} < 1$. Note that this is the same parameterisation as the one used in Gonzalo and Pitarakis (2012) but its key details are repeated here for self containedness considerations. Throughout this paper we will also refer to the true value of the threshold parameter as either γ_0 or λ_0 when relevant.

Our main goal is to focus on the sole predictive power of x_t without imposing any restrictions on the α's. Note for instance that a null hypothesis such as $\alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ may be rejected solely due to $\alpha_1 \neq \alpha_2$ while continuing to be compatible with an environment in which x_t has no predictive content. It is this aspect that we wish to address in the present paper whose goal is to develop inferences about the β's without imposing any constraints on the α's in the sense that they may or may not be regime dependent. More specifically we will be interested in exploring testing strategies for testing the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ while allowing the α's to be free in the background. In what follows we outline our operating assumptions regarding the probabilistic properties of u_t, v_t, q_t and their joint interactions. Throughout this paper we let the random disturbance v_t be described by the linear process $v_t = \Psi(L)e_{vt}$ with the polynomial $\Psi(L) = \sum_{j=0}^{\infty} \Psi_j L^j$ having $\Psi(1) \neq 0$, $\Psi_0 = 1$ and absolutely summable coefficients. We also let $\zeta_t = (u_t, e_{vt})'$ and introduce the filtration $\mathcal{F}_t = \sigma(\zeta_s, u_{qs} | s \leq t)$.

ASSUMPTIONS *A1: $E[\zeta_t|\mathcal{F}_{t-1}] = 0$, $E[\zeta_t \zeta_t'|\mathcal{F}_{t-1}] = \widetilde{\Sigma} > 0$, $\sup_t E\zeta_{it}^4 < \infty$. A2: The sequence $\{u_{qt}\}$ is strictly stationary, ergodic, strong mixing with mixing numbers α_m such that $\sum_{m=1}^{\infty} \alpha^{\frac{1}{m}-\frac{1}{r}} < \infty$ for some $r > 2$. A3: The probability density function $f_q(.)$ of q_t is bounded away from zero and ∞ over each bounded set.*

Assumption A1 requires the error process driving (1) to be a martingale difference sequence with respect to $\mathcal{F}_t$ hence rules out serial correlation in u_t (but not in v_t or q_t) while also imposing conditional homoskedasticity. Both v_t and q_t are allowed to be sufficiently general dependent processes. This setting mimics closely the standard framework used in the predictive regression literature (e.g. Campbell and Yogo (2006), Jansson and Moreira (2006)) and is in fact slightly more general since we do allow v_t to be serially correlated. At this stage it is also important to clarify our stance regarding the joint interactions of our variables. Our assumptions about the dependence structure of the random disturbances together with the finiteness of moments requirements imply that a Functional Central Limit Theorem holds for $w_t = (u_t, u_t I_{1t-1}, v_t)$. More formally $T^{-\frac{1}{2}} \sum_{t=1}^{[Tr]} w_t \Rightarrow (B_u(r), B_u(r, \lambda), B_v(r)' = BM(\Omega)$ with $\Omega = \sum_{k=-\infty}^{\infty} E[w_0 w_k']$. Our analysis will impose a particular structure on $\Omega = [\omega_{ij}]\ i, j = 1, 2, 3$ which governs and restricts the joint interactions of u_t, v_t and q_t. More specifically we impose

$$\Omega = \begin{pmatrix} \sigma_u^2 & \lambda\sigma_u^2 & \sigma_{uv}\Psi(1) \\ \lambda\sigma_u^2 & \lambda\sigma_u^2 & \lambda\sigma_{uv}\Psi(1) \\ \sigma_{uv}\Psi(1) & \lambda\sigma_{uv}\Psi(1) & \sigma_e^2\Psi(1)^2 \end{pmatrix} \tag{3}$$

where $\sigma_u^2 = E[u_t^2]$, $\sigma_v^2 = E[v_t^2]$ and since $E[u_t e_{v,t-j}] = 0$ we also write $\sigma_{uv} = E[u_t v_t] = E[u_t v_t] = \sigma_{ue}$. The chosen structure of Ω is general enough to encompass the standard setting used in the linear predictive regression literature that typically imposes $\{u_t, v_t\}$ to be a martingale difference sequence and u_t and v_t solely contemporaneously correlated. Our assumptions allow us to operate within a similar environment while also permitting the shocks to the threshold variable to be contemporaneously correlated with u_t and/or v_t. As in Caner and Hansen (2001) and Pitarakis (2008), $B_u(r, \lambda)$ refers to a two-parameter Brownian Motion which is a zero mean Gaussian process with covariance kernel $(r_1 \wedge r_2)(\lambda_1 \wedge \lambda_2)\sigma_u^2$ so that we implicitly also operate under the requirement that $E[u_t^2|q_{t-1}, q_{t-2}, \ldots] = \sigma_u^2$ as well as $E[u_t v_t|q_{t-1}] = E[u_t v_t] \equiv \sigma_{uv}$ and $E[u_t v_{t-k}|q_{t-1}, q_{t-2}, \ldots] = 0 \ \forall k \geq 1$. Given our nearly integrated specification for x_t and A1-A3 above it is also clear (see Phillips (1988)) that $x_{[Tr]}/\sqrt{T} \Rightarrow J_c(r)$ with $J_c(r) = B_v(r) + c\int_0^r e^{(r-s)c}B_v(s)ds$ denoting a scalar Ornstein-Uhlenbeck process. For later use we also define the demeaned versions of $J_c(r)$ and $B_u(r)$ as $J_c^*(r) = J_c(r) - \int J_c(r)$ and $B_u^*(r) = B_u(r) - \int B_u(r)$.

3 Large Sample Inference

Since within model (1) $H_0 : \beta_1 = \beta_2 = 0$ is compatible with either $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$, in a first instance it will be important to establish the large sample properties of our threshold parameter estimator $\hat{\gamma}$ (or $\hat{\lambda}$) under the two alternative scenarios on the intercepts.

3.1 Threshold Paramater Estimation

The threshold parameter estimator we consider throughout this paper is based on the least squares principle and defined as

$$\widehat{\gamma} \;\; = \;\; \arg\min_{\gamma} S_T(\gamma) \tag{4}$$

with $S_T(\gamma)$ denoting the concentrated sum of squared errors function obtained from (1) under the restriction $\beta_1 = \beta_2 = 0$ i.e. $S_T(\gamma) = y'y - \sum_{i=1}^2 y'I_i(I_i'I_i)^{-1}I_i'y$. Recall that throughout this paper we use $\hat{\gamma}$ and $\hat{\lambda} = \arg\min_{\lambda \in (0,1)} S_T(\lambda)$ interchangeably. Naturally, the behaviour of $\hat{\lambda}$ is expected to depend on whether the underlying true model has $\alpha_1 \neq \alpha_2$ (i.e. identified threshold parameter) or $\alpha_1 = \alpha_2$ in which case λ vanishes from the true model. The following Proposition summarises the large sample behaviour of $\hat{\lambda}$ under the two scenarios.

Proposition 1. *Under Assumptions A1-A3, $H_0 : \beta_1 = \beta_2 = 0$ and as $T \to \infty$ we have (i) $T|\hat{\lambda} - \lambda_0| = O_p(1)$ when $\alpha_1 \neq \alpha_2$ and (ii) $\hat{\lambda} \xrightarrow{d} \lambda^*$ with $\lambda^* = \arg\max_{\lambda \in \Lambda}[B_u(\lambda) - \lambda B_u(1)]^2/\lambda(1-\lambda)$ when $\alpha_1 = \alpha_2$.*

When $\beta_1 = \beta_2 = 0$ is imposed on the fitted model and $\alpha_1 \neq \alpha_2$ we have a purely stationary mean shift specification and the result in part (i) of Proposition 1 is intuitive and illustrates the T-consistency of the least squares based threshold parameter estimator. This is in fact a well known result in the literature which we report for greater coherence with our subsequent analysis (see Hansen (2000) and Gonzalo and Pitarakis (2002)). The result in part (ii) of Proposition 1 is particularly interesting and highlights the fact that the threshold parameter estimator obtained from a model that is linear and contains no threshold effects converges in distribution to a random variable given by the maximum of a normalised squared Brownian Bridge process. Although the maximum of a Brownian Bridge is well known to be a uniformly distributed random variable an explicit expression or closed form density for λ^* is to our knowledge not available in the literature.

We next concentrate on the limiting distribution of a Wald type test statistic for testing $H_0 : \beta_1 = \beta_2 = 0$ in (1).

3.2 Testing $H_0 : \beta_1 = \beta_2 = 0$

For a given $\lambda \in (0, 1)$ and letting $R = \{(0, 1, 0, 0), (0, 0, 0, 1)\}$ denote the 2×4 restriction matrix we write the Wald statistic for testing $H_0 : \beta_1 = \beta_2 = 0$ in (1) as

$$W_T(\lambda) = \hat{\theta}(\lambda)' R'(R(Z(\lambda)' Z(\lambda))^{-1} R)^{-1} R\hat{\theta}(\lambda)/\hat{\sigma}_u^2(\lambda) \tag{5}$$

with $\hat{\sigma}_u^2(\lambda)$ referring to the residual variance estimated from the unrestricted specification in (1). In what follows $W_T(\hat{\lambda})$ denotes the Wald statistic evaluated at the estimated threshold parameter $\hat{\lambda}$ as defined in (4). The limiting behaviour of $W_T(\hat{\lambda})$ is now summarised in the following Proposition.

Proposition 2 *Under the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$, assumptions A1-A3 and as $T \to \infty$ we have*

$$W_T(\hat{\lambda}) \Rightarrow \frac{\left[\int J_c^*(r) dB_u(r, 1)\right]^2}{\sigma_u^2 \int J_c^*(r)^2} + \chi^2(1) \tag{6}$$

regardless of whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$.

Proposition 2 above highlights the usefulness of the Wald statistic for conducting inferences about the $\beta' s$ without having to take a stand on whether the $\alpha' s$ are regime dependent or not. The interesting point here is the fact that the limiting distribution of the Wald statistic evaluated at $\hat{\lambda}$ is the same regardless of whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$ in the underlying model. One shortcoming of our expression in (6) is caused by the presence of the unknown noncentrality parameter c making it difficult to tabulate in practice. Due to the allowed correlation between B_u and B_v it is also the case that the first component in the right hand side of (6) will depend on σ_{uv}. There is however an instance under which the limiting distribution simplifies considerably as summarised in Proposition 3 below.

Proposition 3 *Under the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$, assumptions A1-A3 together with the requirement that $\sigma_{uv} = 0$ (exogeneity) in (3) and as $T \to \infty$ we have*

$$W_T(\hat{\lambda}) \Rightarrow \chi^2(2) \tag{7}$$

regardless of whether $\alpha_1 = \alpha_2$ *or* $\alpha_1 \neq \alpha_2$.

The above result highlights a unique scenario whereby the magnitude of the noncentrality parameter no longer enters the asymptotics of the Wald statistic despite a nearly integrated parameterisation in the DGP. See also Rossi (2005) for interesting similarities between our asymptotics in Proposition 2 and distributions arising within a related structural break framework.

We next introduce an Instrumental Variable based modified Wald statistic designed in such a way that its limiting distribution remains a nuisance parameter free $\chi^2(2)$ random variable regardless of whether σ_{uv} is zero or not. This is achieved through an IV method developed in Phillips and Magdalinos (2009) in the context of the cointegration literature and which we adapt to our current context (see also Breitung and Demetrescu (2014)). The key idea is to instrument x_t with a slightly less persistent version of itself using its own innovations (hence the IVX terminology). Letting $\phi_T = (1 - c_z/T^{\delta})$ for some $c_z > 0$ (say $c_z = 1$ as discussed in Phillips and Magdalinos (2009) and Kostakis, Magdalinos and Stamatogiannis (2014)) and $\delta \in (0, 1)$ the IV variable is constructed as $\tilde{h}_t = \sum_{j=1}^{t} \phi_T^{t-j} \Delta x_j$. Within our present context we instrument $x_t I_{it}$ in (1) with $\tilde{h}_t I_{it}$ for $i = 1, 2$ and refer to the vectors stacking the $\tilde{h}_t I_{1t}$ and $\tilde{h}_t I_{2t}$ observations as $\tilde{h}_1$ and $\tilde{h}_2$. We also define $\widetilde{H}_1 = (I_1\ \tilde{h}_1)$, $\widetilde{H}_2 = (I_2\ \tilde{h}_2)$ and let $\widetilde{H} = (\widetilde{H}_1\ \widetilde{H}_2)$. The IV based estimator of $\theta(\lambda)$ in (1), say $\hat{\theta}^{IV}(\lambda)$ can now be formulated as

$$\hat{\theta}^{IV}(\lambda) = (\widetilde{H}(\lambda)' Z(\lambda))^{-1} \widetilde{H}(\lambda)' y. \tag{8}$$

and the IV based Wald statistic for testing $\beta_1 = \beta_2 = 0$ in (1) is given by

$$W_T^{IV}(\lambda) = (R\, \hat{\theta}^{IV}(\lambda))' [R (\widetilde{H}(\lambda)' Z(\lambda))^{-1} \widetilde{H}(\lambda)' \widetilde{H}(\lambda) (\widetilde{H}(\lambda)' Z(\lambda))^{-1} R']^{-1} (R\, \hat{\theta}^{IV}(\lambda)) / \hat{\sigma}^2(\lambda). \tag{9}$$

Proposition 4 *Under the null hypothesis* $H_0 : \beta_1 = \beta_2 = 0$*, assumptions A1-A3 and as* $T \to \infty$ *we have* $W_T^{IV}(\hat{\lambda}) \Rightarrow \chi^2(2)$ *regardless of whether* $\alpha_1 = \alpha_2$ *or* $\alpha_1 \neq \alpha_2$.

The above results provides a convenient test statistic for testing $H_0 : \beta_1 = \beta_2 = 0$. Inferences are based on a limiting distribution that does not depend on c or any endogeneity induced parameter (as opposed to our formulation in (6)). The parameter δ controls the degree of persistence of the Instrumental Variable.

4 Finite Sample Evaluation

The goal of this section is twofold. First, we wish to demonstrate the validity and finite sample accuracy of our theoretical results presented in Propositions 3-4 through simulations. Second, we wish to evaluate the finite sample performance of our Wald statistics through size and power experiments. We initially concentrate on our result stating that the limiting distribution of the

Wald statistic for testing $H_0 : \beta_1 = \beta_2 = 0$ in (1) is $\chi^2(2)$ *regardless* of whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$ and regardless of the magnitude of the noncentrality parameter c appearing in the DGP.

Our chosen DGP is given by (1) with $\beta_1 = \beta_2 = 0$. For the parameterisation of the intercepts we consider two scenarios. Namely, $\{\alpha_1, \alpha_2\} = \{1, 1\}$ and $\{\alpha_1, \alpha_2\} = \{1, 3\}$. In the latter case we set $\gamma_0 = 0.25$ with the threshold variable taken to follow the AR(1) process $q_t = 0.5q_{t-1} + u_{qt}$ while we set $v_t = 0.5v_{t-1} + e_t$ for the shocks associated with the nearly integrated variable x_t. Finally we take (u_t, e_t, u_{qt}) to be a Gaussian vector with covariance given by $\Sigma = \{(1, \sigma_{ue}, \sigma_{uq}), (\sigma_{eu}, 1, \sigma_{eq}), (\sigma_{uq}, \sigma_{eq}, 1)\}$. We initially focus on a scenario characterised by exogeneity setting $\Sigma = Id_3$ and subsequently also consider the more general case that allows contemporaneous correlations across all random disturbances by setting $(\sigma_{ue}, \sigma_{uq}, \sigma_{eq}) = (-0.5, 0.3, 0.4)$.

Table 1 below displays the simulated finite sample critical values of $W_T(\hat{\lambda})$ together with those of the $\chi^2(2)$ under $c = 1$ and $c = 10$. The number of Monte-Carlo draws is set at $N = 5000$ throughout. Overall we note an excellent match of the simulated quantiles with their asymptotic counterparts. It is also clear that varying c has little impact on the quantiles as expected by our result in Proposition 3. Perhaps more importantly we note the robustness of the estimated quantiles to the two scenarios about the $\alpha's$. Even under moderately small sample sizes such as $T = 200$, the cutoffs of the asymptotic distribution of $W_T(\hat{\lambda})$ under $\alpha_1 = \alpha_2$ and $\alpha_1 \neq \alpha_2$ remain extremely close as confirmed by our theory.

Table 1. Simulated Quantiles of $W_T(\hat{\lambda})$ versus $\chi^2(2)$ under Exogeneity

	$\alpha_1 = \alpha_2$			$\alpha_1 \neq \alpha_2$		
	10%	5%	2.5%	10%	5%	2.5%
$\chi^2(2)$	4.605	5.991	7.378	4.605	5.991	7.378
	$c = 1$			$c = 1$		
$T = 200$	4.900	6.397	7.833	5.032	6.525	7.800
$T = 400$	4.876	6.225	7.833	4.880	6.265	7.820
$T = 1000$	4.643	5.997	7.396	4.690	6.209	7.443
	$c = 10$			$c = 10$		
$T = 200$	4.851	6.527	8.052	4.929	6.324	7.916
$T = 400$	4.608	6.122	7.509	4.678	6.135	7.666
$T = 1000$	4.635	5.879	7.104	4.673	6.072	7.388

We next, concentrate on our IV based Wald statistic and evaluate the finite sample adequacy of the asymptotic $\chi^2(2)$ approximation to its distribution under endogeneity. We initially evaluate the quantiles of its distribution across alternative scenarios and compare them with those of a $\chi^2(2)$ random variable. We subsequently implement a formal size experiment comparing empirical sizes with their nominal counterparts. Our results on the quantiles of $W_T^{IV}(\hat{\lambda})$ are presented in Table 2. Note that for this set of experiments our IV variables have been generated using the persistence parameter $\delta = 0.7$ while in our subsequent and more formal size experiments we highlight more extensively the sensitivity of the distributional properties of ${W_T}^{IV}(\hat{\lambda})$ to alternative magnitudes of δ (see Table 3).

Focusing on Table 2 we note the close proximity of the quantiles of the distribution of ${W_T}^{IV}(\hat{\lambda})$ to those of the $\chi^2(2)$ regardless of the magnitude of the noncentrality parameter c or whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$. The accurate matching of the quantiles also appears to be maintained across moderately small sample sizes. Under T=400 for instance the 10% estimated quantile of ${W_T}^{IV}(\hat{\lambda})$ was 4.876 under $\alpha_1 = \alpha_2$ and 4.880 under $\alpha_1 \neq \alpha_2$ compared with 4.605 for the theoretical $\chi^2(2)$ counterpart. It is also clear however that for sample sizes such as $T = 200$ the finite sample quantiles are slightly above their asymptotic counterparts which may result in mild size distortions. An issue that is investigated below.

Table 2. Simulated Quantiles of $W_T^{IV}(\hat{\lambda}, \delta = 0.7)$ versus $\chi^2(2)$ under Endogeneity

	$\alpha_1 = \alpha_2$			$\alpha_1 \neq \alpha_2$		
	10%	5%	2.50%	10%	5%	2.50%
$\chi^2(2)$	4.605	5.991	7.378	4.605	5.991	7.378
		c=1			c=1	
$T = 200$	4.900	6.397	7.833	5.032	6.525	7.800
$T = 400$	4.876	6.225	7.833	4.880	6.265	7.820
$T = 1000$	4.643	5.997	7.396	4.690	6.209	7.443
		c=10			c=10	
$T = 200$	4.851	6.527	8.052	4.929	6.324	7.916
$T = 400$	4.608	6.122	7.509	4.678	6.135	7.666
$T = 1000$	4.635	5.879	7.104	4.673	6.072	7.388

We next focus on the implied empirical size properties of $W_T^{IV}(\hat{\lambda})$ by evaluating the number of times the computed pvalue of our test statistic exceeds a given nominal percentage. Results are presented in Table 3 below. One additional goal of the present exercise is to highlight the sensitivity of our IV based method to the choice of δ needed for constructing the instrumental variable.

Table 3. Empirical Size of IV Corrected and Uncorrected Wald Statistics

	$W_T^{IV}(\hat{\lambda})$			$W_T(\hat{\lambda})$			$W_T^{IV}(\hat{\lambda})$			$W_T(\hat{\lambda})$		
	10%	5%	2.5%	10%	5%	2.5%	10%	5%	2.5%	10%	5%	2.5%
T=200, c=1				$\alpha_1 = \alpha_2$						$\alpha_1 \neq \alpha_2$		
$\delta = 0.70$	11.6	6.3	3.1	15.0	8.7	5.0	12.6	6.3	3.4	15.5	8.9	4.7
$\delta = 0.75$	11.8	6.7	3.5	15.0	8.7	5.0	12.7	6.7	3.5	15.5	8.9	4.7
$\delta = 0.80$	12.4	7.0	3.6	15.0	8.7	5.0	13.2	7.1	3.7	15.5	8.9	4.7
$\delta = 0.85$	12.8	7.2	3.9	15.0	8.7	5.0	13.7	7.1	3.8	15.5	8.9	4.7
T=400, c=1												
$\delta = 0.70$	10.9	5.9	3.0	14.2	8.1	4.6	11.3	5.6	3.0	14.4	7.9	4.7
$\delta = 0.75$	11.4	6.4	3.0	14.2	8.1	4.6	11.9	5.2	3.2	14.4	7.9	4.7
$\delta = 0.80$	12.2	6.3	3.3	14.2	8.1	4.6	12.2	6.2	3.4	14.4	7.9	4.7
$\delta = 0.85$	13.0	6.7	3.6	14.2	8.1	4.6	12.5	6.5	3.5	14.4	7.9	4.7
T=1000, c=1												
$\delta = 0.70$	10.2	5.0	2.6	14.8	8.3	4.5	10.5	5.5	2.6	14.4	7.9	4.7
$\delta = 0.75$	10.6	5.2	2.6	14.8	8.3	4.5	11.2	5.5	2.7	14.4	7.9	4.7
$\delta = 0.80$	11.1	5.4	2.7	14.8	8.3	4.5	11.3	5.9	2.8	14.4	7.9	4.7
$\delta = 0.85$	11.4	5.9	2.8	14.8	8.3	4.5	12.0	6.3	3.2	14.4	7.9	4.7
T=200, c=10												
$\delta = 0.70$	11.1	6.2	3.2	12.1	6.7	3.7	11.6	6.1	3.2	12.7	6.9	3.7
$\delta = 0.75$	11.4	6.5	3.2	12.1	6.7	3.7	11.6	6.1	3.3	12.7	6.9	3.7
$\delta = 0.80$	11.6	6.6	3.4	12.1	6.7	3.7	11.5	6.4	3.4	12.7	6.9	3.7
$\delta = 0.85$	11.6	6.6	3.5	12.1	6.7	3.7	11.6	6.5	3.4	12.7	6.9	3.7
T=400, c=10												
$\delta = 0.70$	10.0	5.4	2.7	11.3	6.1	3.3	10.4	5.4	2.9	11.6	6.1	3.4
$\delta = 0.75$	10.7	5.5	2.9	11.3	6.1	3.3	10.2	5.2	2.2	11.6	6.1	3.4
$\delta = 0.80$	10.6	5.3	3.0	11.3	6.1	3.3	10.5	5.4	2.7	11.6	6.1	3.4
$\delta = 0.85$	10.2	4.9	2.6	11.3	6.1	3.3	10.6	5.7	2.7	11.6	6.1	3.4
T=1000, c=10												
$\delta = 0.70$	10.3	4.7	2.2	10.5	5.2	2.5	10.5	5.5	2.6	11.2	5.7	2.8
$\delta = 0.75$	9.90	4.8	2.2	10.5	5.2	2.5	10.3	5.2	2.6	11.2	5.7	2.8
$\delta = 0.80$	9.70	5.0	2.3	10.5	5.2	2.5	11.3	5.9	2.8	11.2	5.7	2.8
$\delta = 0.85$	9.60	4.8	2.4	10.5	5.2	2.5	10.1	5.0	2.8	11.2	5.7	2.8

Comparing the performance of $W_T^{IV}(\hat{\lambda})$ and $W_T(\hat{\lambda})$ (using $\chi^2(2)$ critical values) it is again clear that our IV based statistic significantly improves upon the standard Wald by bringing the implied empirical sizes significantly closer to their nominal counterparts. It is also the case however that across some scenarios $W_T^{IV}(\hat{\lambda})$ may also be subject to mild to moderate size distortions. This happens as the persistence parameter δ approaches 1 leading to $W_T^{IV}(\hat{\lambda})$ being mildly oversized in small to moderate sample sizes. This is perhaps not surprising since our IV variable approaches the original regressor as $\delta \to 1$. Under $\delta = 0.7$ however we note an overall good match of empirical and nominal sizes regardless of whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$. Although the magnitude of the noncentrality parameter c is also of no influence asymptotically, our results in Table 3 suggest a mild improvement of size properties under $c = 10$ versus $c = 1$. In the former case and given the sample sizes the predictor variable is significantly further from the nonstationarity region suggesting that standard inferences should apply.

Finally in our last experiment we document the ability of our test statistic to detect fixed departures from the null hypothesis across a wide range of paramaterisations. Results are presented in Table 4 below.

Table 4. Empirical Power of IV Corrected Wald Statistic (2.5% level)

$\alpha_1 = \alpha_2$	$c = 1, \delta = 0.70$		$c = 1, \delta = 0.75$		$c = 1, \delta = 0.80$	
$\beta_1 = 0$	$\beta_2 = 0.025$	$\beta_2 = 0.05$	$\beta_2 = 0.025$	$\beta_2 = 0.05$	$\beta_2 = 0.025$	$\beta_2 = 0.05$
$T = 200$	0.195	0.610	0.206	0.657	0.248	0.704
$T = 400$	0.570	0.942	0.656	0.964	0.697	0.977
$T = 1000$	0.998	1.000	1.000	1.000	1.000	1.000
	$c = 10, \delta = 0.70$		$c = 10, \delta = 0.75$		$c = 10, \delta = 0.80$	
$T = 200$	0.079	0.266	0.074	0.275	0.083	0.305
$T = 400$	0.262	0.852	0.275	0.882	0.311	0.893
$T = 1000$	0.961	1.000	0.977	1.000	0.980	1.000
$\alpha_1 \neq \alpha_2$	$c = 1, \delta = 0.70$		$c = 1, \delta = 0.75$		$c = 1, \delta = 0.80$	
$\beta_1 = 0$	$\beta_2 = 0.025$	$\beta_2 = 0.05$	$\beta_2 = 0.025$	$\beta_2 = 0.05$	$\beta_2 = 0.025$	$\beta_2 = 0.05$
$T = 200$	0.252	0.686	0.269	0.730	0.309	0.773
$T = 400$	0.656	0.947	0.732	0.966	0.777	0.979
$T = 1000$	0.964	1.000	0.981	1.000	0.992	1.000
	$c = 10, \delta = 0.70$		$c = 10, \delta = 0.75$		$c = 10, \delta = 0.80$	
$T = 200$	0.111	0.397	0.098	0.395	0.111	0.425
$T = 400$	0.383	0.944	0.417	0.966	0.455	0.966
$T = 1000$	0.964	1.000	0.992	1.000	0.998	1.000

As expected from our earlier size experiments, in finite samples power is increasing in δ highlighting an obvious size/power tradeoff when it comes to selecting a suitable δ in practice. Power is clearly increasing with the sample size reaching a probability of correct decision close to 100% under T=1000. For mild deviations from the null hypothesis (e.g. $\beta_1 = 0$ to $\beta_2 = 0.025$) empirical power is at about 20% under T=200 climbing up to 57% with $T = 400$. It is also interesting to highlight the distinct finite sample behaviour of the test statistic when $c = 1$ versus $c = 10$. In the latter case, empirical power is significantly lower unless T is very large which is consistent with our earlier size based results.

5 Valuation Ratio Based Return Predictability

Due to its ability to let the data determine the presence or absence of regime specific behaviour in predictive regressions, our threshold setting is particularly suited for exploring the presence of time varying return predictability when time variation is driven by economic episodes such as recessions and expansions rather than calendar time per se. The new inference theory developed in this paper is an important complement to the two test statistics proposed in Gonzalo and Pitarakis (2012) allowing us to distinguish between regime specific predictability truly induced by a particular predictor such as the Dividend Yield and regime specific behaviour that may arise solely due to the variable used for generating the regimes (e.g. a business cycle proxy).

Despite a huge literature geared towards testing for the linear predictability of stock returns with valuation ratios such as the Dividend Yield, it is only recently that empirical work has recognised the possibility that predictability may be kicking in occasionally depending on the state of the economy. In Gonzalo and Pitarakis (2012) for instance, using aggregate US data over the 1950-2007 period we established a strong countercyclical property to Dividend Yield based predictability of stock returns with an R^2 as high as 17% in the weak or negative growth regime dropping to 0% during expansions (see also Henkel, Martin and Nardari (2011) who reached similar conclusions using a different statistical framework). More recently Gargano (2013) also reached similar conclusions using the Dividend to Price ratio as a predictor while also proposing a theoretical framework that embeds this recessionary period based predictability of stock returns within a consumption based asset pricing model. Earlier research that highlighted the importance of a changing environment on predictability include Pesaran and Timmermann (1995), Paye and Timmermann (2006) amongst numerous others.

We here consider the question of episodic predictability of aggregate US market returns using the four most commonly considered valuation ratios, namely the Dividend Yield (DY), Book-to-Market ratio (BM), the Dividend to Price ratio (DP) and the Earnings Yield (EP) all expressed in natural logs. Our predictability episodes are driven by the monthly growth rate in the US industrial production index used as a proxy for the state of the economy. Compared to our analysis in Gonzalo and Pitarakis (2012) where we had focused solely on DY we also extend our sample to cover the 1927-2013 period using the recently extended Goyal and Welch data set (see Goyal and Welch (2014) and Welch and Goyal (2008)). The specific return series we are considering is the recently revised excess returns series referred to as $Mkt - RF$ in Kenneth French's data library with Mkt referring to the value weighted returns of all CRSP firms listed on the NYSE, AMEX or NASDAQ and RF the one month T-Bill return.

Before proceeding with the above analysis it is important to reconsider our findings in Gonzalo and Pitarakis (2012) where we had explored the predictive power of DY over the 1950-2007 period using a slightly different definition of aggregate market returns. There, we had documented a very strong ability of the DY to predict returns during bad times or recessions. The null hypotheses $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ and $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ were rejected on the basis of computed test statistics given by $SupA = 20.75$ [0.001] and $SupB^{ivx}(\delta = 0.7) = 26.75$ [0.000]. Using the same data as in Gonzalo and Pitarakis (2012), inferences based on test statistic introduced in this paper lead to $W_T(\hat{\lambda}, \delta = 0.7) = 6.795$ [0.033], $W_T(\hat{\lambda}, \delta = 0.8) = 8.619$ [0.013] and $W_T(\hat{\lambda}, \delta = 0.9) = 9.453$ [0.009] further corroborating our claim of regime specific predictability induced by the Dividend Yield itself.

Next, focusing on the new set of predictors our key results are displayed in Table 5 where we present the magnitude of our test statistics across alternative choices of the persistence parameter δ used in the construction of the IVX variable. Figures in square brackets are pvalues. We recall that the SupA and $SupB^{ivx}$ test statistics are associated with the null hypotheses given by $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ and $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ while $W_T(\hat{\lambda}, \delta)$ is designed to test $H_0 : \beta_1 = \beta_2 = 0$. The symbol *** indicates rejection at 2.5% or below, ** at 5% and * at 10%.

Table 5. Regime Specific Predictability of Valuation Ratios

	SupA	SupB			$W_T(\hat{\lambda}, \delta)$		
		$\delta = 0.7$	$\delta = 0.8$	$\delta = 0.9$	$\delta = 0.7$	$\delta = 0.8$	$\delta = 0.9$
				1927-2013			
DY	27.543 [0.000]	33.974***	35.182***	35.076***	7.555 [0.023]	8.158 [0.017]	7.392 [0.025]
BM	34.721 [0.000]	41.187***	41.764***	41.862***	6.227 [0.044]	6.292 [0.043]	6.220 [0.045]
DP	19.193 [0.002]	22.726***	24.019***	24.262***	5.110 [0.078]	5.940 [0.054]	5.544 [0.063]
EP	11.428 [0.065]	14.562*	15.21**	15.528**	3.185 [0.203]	3.776 [0.151]	3.977 [0.137]
				1940-2013			
DY	20.017 [0.002]	24.661***	26.406***	26.945***	4.772 [0.092]	7.174 [0.028]	8.251 [0.016]
BM	11.456 [0.065]	12.617*	13.154*	13.388*	0.961 [0.619]	1.640 [0.440]	2.047 [0.359]
DP	19.366 [0.002]	22.772***	24.513***	25.225***	3.395 [0.183]	5.718 [0.057]	6.983 [0.030]
EP	2.594 [0.974]	3.684	4.458	4.946	1.272 [0.529]	1.944 [0.378]	2.405 [0.300]
				1950-2013			
DY	21.532 [0.001]	24.139***	26.015***	27.795***	3.185 [0.203]	4.817 [0.090]	6.120 [0.047]
BM	12.102 [0.050]	12.256*	12.488*	12.761*	2.542 [0.281]	1.811 [0.404]	1.603 [0.449]
DP	20.227 [0.001]	21.971***	23.654***	25.278***	2.177 [0.337]	3.697 [0.157]	4.908 [0.086]
EP	3.211 [0.911]	3.352	3.717	4.335	0.568 [0.753]	0.790 [0.674]	1.299 [0.522]
				1960-2013			
DY	19.604 [0.002]	21.813***	21.725***	21.594***	2.115 [0.347]	2.348 [0.309]	2.476 [0.290]
BM	10.877 [0.082]	10.92	10.919	10.933	0.043 [0.979]	0.020 [0.990]	0.095 [0.953]
DP	18.233 [0.003]	19.658***	19.808***	19.789***	1.233 [0.540]	1.627 [0.443]	1.846 [0.397]
EP	1.188 [1.000]	1.432	1.458	1.502	0.655 [0.721]	0.475 [0.789]	0.399 [0.819]

Focusing first on the Dividend Yield series we note a consistent and strong rejection of the SupA and $SupB^{ivx}$ based null hypotheses throughout the full sample and the three subperiods.

This further corroborates and strengthens our findings in Gonzalo and Pitarakis (2012) where we had documented the countercyclical predictability of DY over the 1950-2007 period. More importantly we here note that our new test statistic also leads to strong rejections of the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$, indicating that predictability is truly driven by the DY predictor rather than unequal intercepts arising from our business cycle proxy. However, it is interesting to note that the new inferences developed in this paper attribute a more ambiguous role to the Dividend Yield as a predictor when considering post 60s samples. Although SupA and $SupB^{ivx}$ based inferences continue to point towards threshold predictability $H_0 : \beta_1 = \beta_2 = 0$ can no longer be rejected on the basis of our $W_T(\hat{\lambda})$ test statistic when considering the 1960-2013 period. This suggests that over this subperiod, SupA and $SupB^{ivx}$ may in fact be rejecting their respective null hypothesis $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ and $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ mainly due to unequal intercepts i.e. the regime specific nature of return predictability may in fact be driven by our business cycle proxy rather than the DY predictor playing a distinct role across expansions versus recessions. This finding also highlights the crucial importance that needs to be given to the time varying nature of predictability when evaluating the predictive power of any variable for future stock returns. Our results are also in line with a recent branch of the predictability literature which argues that DY based predictability has declined due to greater dividend smoothing. Finally, one may also conjecture that results for the later subperiods may be less reliable due to the significant drop in degrees of freedom. Having estimated a DY based threshold specification for each subsample however we obtained very similar magnitudes for $\hat{\gamma}$ and regime proportions that varied little across the four scenarios of Table 5 i.e. $\{(25\%, 75\%), (22\%, 78\%), (20\%, 80\%), (18\%, 82\%)\}$, ensuring a reasonably large number of observations in each regime.

Regarding the Book-to-Market predictor, it is here interesting to note that at a 4% level or below our new test statistic is unable to reject the null $H_0 : \beta_1 = \beta_2 = 0$ across all scenarios while both SupA and $SupB^{ivx}$ strongly reject their respective null hypotheses when the full sample is considered. This again suggests that any indication of predictability induced by BM may in fact be driven by unequal intercepts rather than the predictive power of BM per se.

For the Dividend-to-Price series and regardless of the sample period considered we note a consistent and strong rejection of the null hypotheses on the basis of our SupA and $SupB^{ivx}$ statistics, indicating strong regime specific effects in the behaviour of stock returns. However in this instance and unlike the DY series our $W_T(\hat{\lambda}, \delta)$ test statistic mostly fails to reject the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$. Although there is some lack of robustness to this result when it comes to experimenting across alternative magnitudes of the IV parameter δ it is quite clear that over the post war period the evidence of any predictive role for DP is weak at best. Finally, across virtually all scenarios it is clear that EP does not contain any predictive power for future excess returns whether linear or regime specific.

6 Conclusions

We developed a toolkit for assessing the predictability induced by a single persistent predictor in an environment that allows predictability to kick in during particular economic episodes and affect all parameters of the model. Our threshold based framework and testing methodology can be used to explore the possibility that the predictive power of highly persistent predictors such as interest rates, valuation ratios and numerous other economic and financial variables may be varying across time in an economically meaningful way with alternating periods of strong versus weak or no predictability. More importantly the core contribution of this paper was to provide a setting that allows us to distinguish predictability induced by a specific predictor from predictability that may be solely driven by economic episodes (e.g. stock returns differing across recessions and expansions). Our empirical results have highlighted the misleading or at best incomplete conclusions one may reach if such regime specific effects are ignored when assessing predictability.

APPENDIX

PROOF OF PROPOSITION 1: Since under $H_0 : \beta_1 = \beta_2 = 0$ the threshold model is given by $y_t = \alpha_1 I_{1t-1} + \alpha_2 I_{2t-1} + u_t$, all assumptions of Gonzalo and Pitarakis (2002) are satisfied implying the statement in (i). The result in Part (ii) follows by first noting that the minimiser of $S_T(\lambda)$ is numerically identical to the maximiser of the Wald statistic $W_T(\lambda)$ for testing $H_0 : \alpha_1 = \alpha_2$ in the above restricted specification. This Wald statistic is given by

$$W_T(\lambda) \quad = \quad \left(\frac{\sum u_t I_{1t-1}}{\sum I_{1t-1}} - \frac{\sum u_t I_{2t-1}}{\sum I_{2t-1}}\right)^2 \frac{\sum I_{1t-1} \sum I_{2t-1}}{T \hat{\sigma}_u^2(\lambda)} \tag{10}$$

with $\hat{\sigma}_u^2(\lambda)$ denoting the residual variance obtained from the above mean shift specification. Under $H_0 : \alpha_1 = \alpha_2$ and A1-A3 a suitable Law of Large Numbers (see White (2000, p.58)) ensures that $\hat{\sigma}_u^2(\lambda) \xrightarrow{p} \sigma_u^2$. From Caner and Hansen (2001) we have $\sum_{t=1}^T u_t I_{1t-1}/\sqrt{T} \Rightarrow B_u(\lambda)$. The strict stationarity and ergodicity of the $I_{it}'s$ further ensures that $\sum I_{1t-1}/T \xrightarrow{p} \lambda$ and $\sum I_{2t-1}/T \xrightarrow{p} (1 - \lambda)$. It now follows from the Continuous Mapping Theorem that

$$W_T(\lambda) \quad \Rightarrow \quad \frac{[B_u(\lambda) - \lambda B_u(1)]^2}{\sigma_u^2 \lambda(1 - \lambda)}. \tag{11}$$

The desired result then follows from the continuity of the argmax functional and the fact that the limit process has a unique maximum in Λ with probability 1 (see Theorem 2.7 in Kim and Pollard (1990)). ■

Before proceeding with the limiting properties of $W_T^{IV}(\hat{\lambda})$ it is useful to recall that in the context of our DGP in (1) standard algebra leads to

$$T\ R\ \hat{\theta}(\lambda) \quad = \quad \begin{pmatrix} \dfrac{\frac{\sum I_{1t-1}}{T} \frac{\sum x_{t-1} y_t I_{1t-1}}{T} - \frac{\sum y_t I_{1t-1}}{\sqrt{T}} \frac{\sum x_{t-1} I_{1t-1}}{T\sqrt{T}}}{\frac{\sum I_{1t-1}}{T} \frac{\sum x_{t-1}^2 I_{1t-1}}{T^2} - \left(\frac{\sum x_{t-1} I_{1t-1}}{T\sqrt{T}}\right)^2} \\ \\ \dfrac{\frac{\sum I_{2t-1}}{T} \frac{\sum x_{t-1} y_t I_{2t-1}}{T} - \frac{\sum y_t I_{2t-1}}{\sqrt{T}} \frac{\sum x_{t-1} I_{2t-1}}{T\sqrt{T}}}{\frac{\sum I_{2t-1}}{T} \frac{\sum x_{t-1}^2 I_{2t-1}}{T^2} - \left(\frac{\sum x_{t-1} I_{2t-1}}{T\sqrt{T}}\right)^2} \end{pmatrix} \equiv \begin{pmatrix} \frac{g_{1t}(\lambda)}{\Delta_{1t}(\lambda)} \\ \\ \frac{g_{2t}(\lambda)}{\Delta_{2t}(\lambda)} \end{pmatrix} \tag{12}$$

and

$$T^2\ R(Z(\lambda)'Z(\lambda))^{-1} R' \quad = \quad \begin{pmatrix} \frac{\sum I_{1t-1}/T}{\Delta_{1t}(\lambda)} & 0 \\ 0 & \frac{\sum I_{2t-1}/T}{\Delta_{2t}(\lambda)} \end{pmatrix}. \tag{13}$$

Given our null hypothesis of interest $H_0 : \beta_1 = \beta_2 = 0$, it is also useful to specialise (12) across the two scenarios on the $\alpha's$, namely $y_t = \alpha + u_t$ if $\alpha_1 = \alpha_2$ and $y_t = \alpha_1 I_{1t-1}^0 + \alpha_2 I_{2t-1}^0 + u_t$ if $\alpha_1 \neq \alpha_2$. In this latter case the quantities I_{1t-1}^0 and I_{2t-1}^0 refer to the indicator functions evaluated at the true threshold parameter λ_0. We write

$$[T\ R\ \hat{\theta}(\lambda)]_{\alpha_1 = \alpha_2} \quad = \quad \begin{pmatrix} \frac{g_{1t}(\lambda)|_{\alpha_1 = \alpha_2}}{\Delta_{1t}(\lambda)} \\ \\ \frac{g_{2t}(\lambda)|_{\alpha_1 = \alpha_2}}{\Delta_{2t}(\lambda)} \end{pmatrix} \tag{14}$$

where

$$g_{it}(\lambda)|_{\alpha_1 = \alpha_2} \quad = \quad \frac{\sum I_{it-1}}{T} \frac{\sum x_{t-1} u_t I_{it-1}}{T} - \frac{\sum u_t I_{it-1}}{\sqrt{T}} \frac{\sum x_{t-1} I_{it-1}}{T\sqrt{T}} \tag{15}$$

for $i = 1, 2$. Similarly,

$$[T\ R\ \hat{\theta}(\lambda)]_{\alpha_1 \neq \alpha_2} \quad = \quad \begin{pmatrix} \dfrac{g_{1t}(\lambda)|_{\alpha_1 \neq \alpha_2}}{\Delta_{1t}(\lambda)} \\ \\ \dfrac{g_{2t}(\lambda)|_{\alpha_1 \neq \alpha_2}}{\Delta_{2t}(\lambda)} \end{pmatrix} \tag{16}$$

with

$$g_{it}(\lambda)|_{\alpha_1\neq\alpha_2} = \frac{\sum I_{it-1}}{T}\left(\alpha_1\frac{\sum x_{t-1}I_{it-1}I^0_{1t-1}}{T}+\alpha_2\frac{\sum x_{t-1}I_{it-1}I^0_{2t-1}}{T}+\frac{\sum x_{t-1}u_tI_{it-1}}{T}\right)-$$

$$\left(\alpha_1\frac{\sum I_{it-1}I^0_{1t-1}}{\sqrt{T}}+\alpha_2\frac{\sum I_{it-1}I^0_{2t-1}}{\sqrt{T}}+\frac{\sum u_tI_{it-1}}{\sqrt{T}}\right)\frac{\sum x_{t-1}I_{it-1}}{T\sqrt{T}}. \tag{17}$$

Before proceeding with the proof of Proposition 2 we introduce the following auxiliary Lemma that is used for establishing the asymptotic properties of the sample moments in (17).

LEMMA A1. Under Assumptions A1-A3, $T|\hat{\lambda}-\lambda_0| = O_p(1)$ and letting $U_t \equiv F(q_t)$, as $T\to\infty$ we have

$$\frac{1}{\sqrt{T}}\sum I(U_{t-1}\leq\hat{\lambda})I(U_{t-1}\leq\lambda_0) - \frac{1}{\sqrt{T}}\sum I(U_{t-1}\leq\lambda_0) \quad \xrightarrow{p} \quad 0 \tag{18}$$

PROOF of LEMMA A1: We need to establish that for every $\varepsilon>0$ and $\delta>0$

$$\lim_{T\to\infty} P\left[\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[I\left(q_t<\widehat{\lambda}\right)-I\left(q_t<\lambda\right)\right]I\left(q_t<\lambda\right)\right|>\varepsilon\right]<\delta.$$

Given that

$$\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[I\left(q_t<\widehat{\lambda}\right)-I\left(q_t<\lambda\right)\right]I\left(q_t<\lambda\right)\right| \leq \left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[I\left(\lambda-\left|\widehat{\lambda}-\lambda\right|<q_t<\lambda+\left|\widehat{\lambda}-\lambda\right|\right)\right]\right|$$
$$\leq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right)$$

with $A_t(\lambda,d) = I(\lambda-|d|<q_t<\lambda+|d|)$, it will be enough to prove that

$$\lim_{T\to\infty} P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right)>\varepsilon\right]<\delta$$

for every $\varepsilon>0$ y $\delta>0$. Since $\hat{\lambda}$ is such that $T|\hat{\lambda}-\lambda_0| = O_p(1)$, therefore for every $\delta>0$, $\exists\Delta_\delta<\infty$ and an integer $T_\delta\geq1$ such that

$$P\left[\left|\widehat{\lambda}-\lambda\right|>\frac{\Delta_\delta}{T}\right]<\delta \qquad \text{for } \forall T>T_\delta,$$

and also

$$P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right)>\varepsilon\right] = P\left[\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right)>\varepsilon\right\}\cap\left\{\left|\widehat{\lambda}-\lambda\right|\leq\frac{\Delta_\delta}{T}\right\}\right]+$$
$$+P\left[\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right)>\varepsilon\right\}\cap\left\{\left|\widehat{\lambda}-\lambda\right|>\frac{\Delta_\delta}{T}\right\}\right]$$
$$\leq P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)>\varepsilon\right]+P\left[\left|\widehat{\lambda}-\lambda\right|>\frac{\Delta_\delta}{T}\right].$$

Using Markov's inequality

$$P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)>\varepsilon\right]\leq\frac{\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)\right\|_1}{\varepsilon}\leq\frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\|A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)\right\|_1}{\varepsilon}$$

and under our assumption on the boundedness of the pdf of q_t away from 0 and ∞ over each bounded set

$$\left\|A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)\right\|_1 = \left\|I\left(\lambda-\frac{\Delta_\delta}{T}<q_t<\lambda+\frac{\Delta_\delta}{T}\right)\right\|_1\leq M\frac{\Delta_\delta}{T}$$

therefore

$$P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)>\varepsilon\right]\leq\frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\|A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)\right\|_1}{\varepsilon}$$
$$\leq\frac{\sqrt{T}M\frac{\Delta_\delta}{T}}{\varepsilon}\leq\frac{M\Delta_\delta}{\varepsilon\sqrt{T}}.$$

Putting together these results we have that for every $\varepsilon>0$ and $\delta>0$ $\exists T_{\varepsilon\delta}<\infty$ such that for every $T>T_{\varepsilon\delta}$

$$P\left[\left|\widehat{\lambda}-\lambda\right|>\frac{\Delta_\delta}{T}\right]<\delta$$
$$P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)>\varepsilon\right]<\delta$$

and then

$$\lim_{T\to\infty}P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right)>\varepsilon\right]\leq\lim_{T\to\infty}P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)>\varepsilon\right]+\lim_{T\to\infty}P\left[\left|\widehat{\lambda}-\lambda\right|>\frac{\Delta_\delta}{T}\right]<2\delta.$$

leading to the desired result. ■

PROOF OF PROPOSITION 2. We initially consider the case $\alpha_1\neq\alpha_2$. Given the T-consistency of $\hat{\lambda}$ for λ_0, $T|\hat{\lambda}-\lambda_0|=O_p(1)$, and our result in Lemma A1 we have

$$g_{it}(\hat{\lambda})|_{\alpha_1\neq\alpha_2}=\frac{\sum I^0_{it-1}}{T}\frac{\sum x_{t-1}u_tI^0_{it-1}}{T}-\frac{\sum u_tI^0_{it-1}}{\sqrt{T}}\frac{\sum x_{t-1}I^0_{it-1}}{T\sqrt{T}}+o_p(1), \tag{19}$$
$$\Delta_{it}(\hat{\lambda})=\frac{\sum I^0_{it-1}}{T}\frac{\sum x^2_{t-1}I^0_{it-1}}{T^2}-\left(\frac{\sum x_{t-1}I^0_{it-1}}{T\sqrt{T}}\right)^2+o_p(1). \tag{20}$$

Using Lemma 1 in Gonzalo and Pitarakis (2012), Theorem 1 in Caner and Hansen (2001) together with the continuous mapping theorem we have

$$\begin{aligned}
g_{1t}(\hat{\lambda})|_{\alpha_1\neq\alpha_2} &\Rightarrow \lambda_0\left(\int J_c(r)dB_u(r,\lambda_0)-B_u(\lambda_0)\int J_c(r)\right),\\
g_{2t}(\hat{\lambda})|_{\alpha_1\neq\alpha_2} &\Rightarrow (1-\lambda_0)\left(\int J_c(r)(dB_u(r)-dB_u(r,\lambda_0))-(B_u(1)-B_u(\lambda_0))\int J_c(r)\right),\\
\Delta_{1t}(\hat{\lambda}) &\Rightarrow \lambda_0^2\int J_c^*(r)^2,\\
\Delta_{2t}(\hat{\lambda}) &\Rightarrow (1-\lambda_0)^2\int J_c^*(r)^2.
\end{aligned} \tag{21}$$

Next, using (20) in (12)-(13) and rearranging gives

$$[T^2R(Z(\hat{\lambda})'Z(\hat{\lambda}))R']^{-1}\Rightarrow\int J_c^*(r)^2\begin{pmatrix}\lambda_0 & 0\\ 0 & (1-\lambda_0)\end{pmatrix} \tag{22}$$

and

$$[TR\hat{\theta}(\hat{\lambda})]_{\alpha_1\neq\alpha_2}\Rightarrow\left(\begin{array}{c}\dfrac{\int J_c(r)dB_u(r,\lambda_0)-B_u(\lambda_0)\int J_c(r)}{\lambda_0\int J_c^*(r)^2dr}\\ \\ \dfrac{\int J_c(r)(dB_u(r)-dB_u(r,\lambda_0))-(B_u(1)-B_u(\lambda_0))\int J_c(r)}{(1-\lambda_0)\int J_c^*(r)^2dr}\end{array}\right). \tag{23}$$

Combining (21)-(22) into (5) and using $\hat{\sigma}^2(\hat{\lambda})\xrightarrow{p}\sigma_u^2$ leads to

$$\begin{aligned}
W_T(\hat{\lambda}) &\Rightarrow \frac{[\int J_cdB_u(r,\lambda_0)-B_u(\lambda_0)\int J_c(r)]^2}{\sigma_u^2\lambda_0\int J_c^*(r)^2}+\\
&\quad\frac{[\int J_c(dB_u(r)-dB_u(r,\lambda_0))-(B_u(1)-B_u(\lambda_0))\int J_c(r)]^2}{\sigma_u^2(1-\lambda_0)\int J_c^*(r)^2}\\
&\equiv \frac{[\int J_c^*(r)dG_u(r,\lambda_0)]^2}{\sigma_u^2\lambda_0(1-\lambda_0)\int J_c^*(r)^2}+\frac{[\int J_c^*(r)dB_u(r)]^2}{\sigma_u^2\int J_c^*(r)^2}\\
&\equiv \frac{[B_u(\lambda_0)-\lambda_0B_u(1)]^2}{\sigma_u^2\lambda_0(1-\lambda_0)}+\frac{[\int J_c^*(r)dB_u(r)]^2}{\sigma_u^2\int J_c^*(r)^2}
\end{aligned} \tag{24}$$

with $G_u(r,\lambda_0) = B_u(r,\lambda_0) - \lambda_0 B_u(r,1)$ denoting a Kiefer Process with covariance function $\sigma_u^2(r_1 \wedge r_2)\lambda_0(1-\lambda_0)$. The result in Proposition 2 then follows by noting that $J_c(r)$ and $G_u(r,\lambda_0)$ are uncorrelated and hence independent due to their Gaussianity so that $\int J_c^*(r)dG_u(r,\lambda) \equiv N(0,\sigma_u^2\lambda_0(1-\lambda_0)\int J_c^*(r)^2)$ conditionally on the realisation of $J_c(r)$. Thus normalising by $\sigma_u^2\lambda_0(1-\lambda_0)\int J_c^*(r)^2$ gives the $\chi^2(1)$ limit which is also the unconditional distribution since not dependent on the realisation of $J_c(r)$. The case $\alpha_1 = \alpha_2$ can be treated in a similar fashion with λ_0 replaced by the random variable λ^*. Indeed, we know that the squared normalised brownian bridge, say X^2, is such that $X = X_\lambda = N(0,1)$ for all given λ weights and the distribution of X is invariant to λ. Therefore this distribution will be maintained when we use $\lambda = \lambda^*$, in spite of the endogeneity between λ^* and the primitive Brownian Motion. ■

PROOF OF PROPOSITION 3. The result follows directly from the independence of $B_u(r,\lambda)$ and $B_v(r)$ under $\sigma_{uv} = 0$ also implying the independence of $J_c^*(r)$ and $B_u(r,\lambda)$ and from which mixed normality follows. ■

Before proceeding with the proof of Proposition 4 it will be convenient to reformulate the components of (9) in an explicit and suitably normalised form. We write

$$T^{1+\delta}R(\widetilde{H}(\lambda)'Z(\lambda))^{-1}\widetilde{H}(\lambda)'\widetilde{H}(\lambda)(\widetilde{H}(\lambda)'Z(\lambda))^{-1}R' = \begin{pmatrix} \dfrac{m_{1t}(\lambda)}{\pi_{1t}(\lambda)} & 0 \\ 0 & \dfrac{m_{2t}(\lambda)}{\pi_{2t}(\lambda)} \end{pmatrix} \tag{25}$$

with

$$\begin{aligned} m_{it}(\lambda) &= \frac{\sum I_{it-1}}{T}\left[\left(\frac{\sum I_{it-1}}{T}\frac{\sum \tilde{h}_{t-1}^2 I_{it-1}}{T^{1+\delta}}\right) - \frac{1}{T^{1-\delta}}\left(\frac{\sum \tilde{h}_{t-1}I_{it-1}}{T^{\frac{1}{2}+\delta}}\right)^2\right] \\ \pi_{it}(\lambda) &= \left(\frac{\sum I_{it-1}}{T}\frac{\sum \tilde{h}_{t-1}x_{t-1}I_{it-1}}{T^{1+\delta}} - \frac{\sum \tilde{h}_{t-1}I_{it-1}}{T^{\frac{1}{2}+\delta}}\frac{\sum x_{t-1}I_{it-1}}{T\sqrt{T}}\right)^2 \end{aligned} \tag{26}$$

for $i = 1, 2$ and

$$R\hat{\theta}^{IV}(\lambda) = \begin{pmatrix} \dfrac{\sum I_{1t-1}\sum y_t\tilde{h}_{t-1}I_{1t-1} - \sum \tilde{h}_{t-1}I_{1t-1}\sum y_t I_{1t-1}}{\sum I_{1t-1}\sum \tilde{h}_{t-1}x_{t-1}I_{1t-1} - \sum \tilde{h}_{t-1}I_{1t-1}\sum x_{t-1}I_{1t-1}} \\ \\ \dfrac{\sum I_{2t-1}\sum y_t\tilde{h}_{t-1}I_{2t-1} - \sum \tilde{h}_{t-1}I_{2t-1}\sum y_t I_{2t-1}}{\sum I_{2t-1}\sum \tilde{h}_{t-1}x_{t-1}I_{2t-1} - \sum \tilde{h}_{t-1}I_{2t-1}\sum x_{t-1}I_{2t-1}} \end{pmatrix}. \tag{27}$$

It will also be useful to rearrange and normalise (27) as follows

$$T^{\frac{\delta+1}{2}}R\hat{\theta}^{IV}(\lambda) = \begin{pmatrix} \dfrac{n_{1t}(\lambda)}{\sqrt{\pi_{1t}(\lambda)}} \\ \\ \dfrac{n_{2t}(\lambda)}{\sqrt{\pi_{2t}(\lambda)}} \end{pmatrix} \tag{28}$$

with

$$n_{it}(\lambda) = \frac{\sum I_{it-1}}{T}\frac{\sum y_t\tilde{h}_{t-1}I_{it-1}}{T^{\frac{1}{2}+\frac{\delta}{2}}} - \frac{1}{T^{\frac{1}{2}+\frac{\delta}{2}}}\left(\frac{\sum \tilde{h}_{t-1}I_{it-1}}{T^{\frac{1}{2}+\delta}}\frac{\sum y_t I_{it-1}}{\sqrt{T}}\right) \tag{29}$$

for $i = 1, 2$.

PROOF OF PROPOSITION 4. We concentrate on the case $\alpha_1 \neq \alpha_2$ with the underlying T-consistency of $\hat{\lambda}$ for λ_0. We also recall that $\tilde{h}_t = \sum_{j=1}^t \phi_T^{t-j}\Delta x_j$ and let $h_t = \sum_{j=1}^t \phi_T^{t-j}v_j$. It now follows directly from (26) and Lemma 3.1 in Phillips and Magdalinos (2009) that

$$\begin{aligned} m_{it}(\hat{\lambda}) &= \left(\frac{\sum I_{it-1}^0}{T}\right)^2\frac{\sum h_{t-1}^2 I_{it-1}^0}{T^{1+\delta}} + o_p(1) \\ \pi_{it}(\hat{\lambda}) &= \left(\frac{\sum h_{t-1}I_{it-1}^0}{T^{\frac{1}{2}+\delta}}\frac{\sum x_{t-1}I_{it-1}^0}{T\sqrt{T}} - \frac{\sum I_{it-1}^0}{T}\frac{\sum h_{t-1}x_{t-1}I_{it-1}^0}{T^{1+\delta}}\right)^2 + o_p(1). \end{aligned} \tag{30}$$

Under our assumptions A1-A3 the following deduce directly from Phillips and Magdalinos (2009, eq. (14))

$$\begin{aligned} m_{1t}(\hat{\lambda}) &\Rightarrow \lambda_0^3 \frac{\omega_v^2}{2} \\ m_{2t}(\hat{\lambda}) &\Rightarrow (1-\lambda_0)^3 \frac{\omega_v^2}{2} \end{aligned} \tag{31}$$

since $\sum h_{t-1}^2 (I_{1t-1}^0 - \lambda_0)/T^{1+\delta} \xrightarrow{p} 0$. It also follows that

$$\begin{aligned} \pi_{1t}(\hat{\lambda}) &\Rightarrow \lambda_0^4 \left[\omega_v^2 + \int J_c^*(r) dJ_c(r)\right]^2 \\ \pi_{2t}(\hat{\lambda}) &\Rightarrow (1-\lambda_0)^4 \left[\omega_v^2 + \int J_c^*(r) dJ_c(r)\right]^2 \end{aligned} \tag{32}$$

so that

$$T^{1+\delta} R(\widetilde{H}(\hat{\lambda})' Z(\hat{\lambda})^{-1} \widetilde{H}(\hat{\lambda})' \widetilde{H}(\hat{\lambda}) (\widetilde{H}(\hat{\lambda})' Z(\hat{\lambda}))^{-1} R' \Rightarrow \begin{pmatrix} \frac{\omega_v^2}{2\lambda_0 \left[\omega_v^2 + \int J_c^*(r) dJ_c(r)\right]^2} & 0 \\ 0 & \frac{\omega_v^2}{2(1-\lambda_0)\left[\omega_v^2 + \int J_c^*(r) dJ_c(r)\right]^2} \end{pmatrix}. \tag{33}$$

Next, we also have

$$n_{it}(\hat{\lambda}) = \frac{\sum I_{it-1}^0}{T} \frac{\sum u_t h_{t-1} I_{it-1}^0}{T^{\frac{1}{2}+\frac{\delta}{2}}} + o_p(1) \tag{34}$$

and Lemma 3.2 in Phillips and Magdalinos (2009) together with (31) ensure the following holds

$$\begin{aligned} \frac{1}{T^{\frac{1}{2}+\frac{\delta}{2}}} \sum h_{t-1} u_t I_{1t-1}^0 &\Rightarrow N(0, \lambda_0^2 \sigma_u^2 \frac{\omega_v^2}{2}) \\ \frac{1}{T^{\frac{1}{2}+\frac{\delta}{2}}} \sum h_{t-1} u_t I_{2t-1}^0 &\Rightarrow N(0, (1-\lambda_0)^2 \sigma_u^2 \frac{\omega_v^2}{2}) \end{aligned} \tag{35}$$

which when rearranged with (33) and using the continuous mapping theorem within $W_T^{IV}(\hat{\lambda})$ leads to the desired result. The case $\alpha_1 = \alpha_2$ can be treated in a similar fashion with λ_0 replaced by the random variable λ^* as formulated in Proposition 1. ■

REFERENCES

Ang, A. and G. Bekaert (2008), "Stock Return Predictability: Is it There?", *Review of Financial Studies*, 20,651-707.

Bandi, F. and B. Perron (2008), "Long Run Risk Return Trade-Offs," *Journal of Econometrics*, 143, 349-374.

Breitung, J. and M. Demetrescu (2014), "Instrumental Variable and Variable Addition Based Inference in Predictive Regressions," *Journal of Econometrics*, forthcoming.

Campbell, J. Y., and Yogo, M. (2006), "Efficient tests of stock return predictability," *Journal of Financial Economics,* 81, 27-60.

Caner, M., and Hansen, B. E. (2001), "Threshold Autoregression with a Unit Root," *Econometrica,* 69, 1555-1596.

Gargano, A., (2013), "Return Predictability in Recessions: An Asset Pricing Perspective," Unpublished Manuscript.

Gonzalo, J., and Pitarakis, J. (2002), "Estimation and Model Selection Based Inference in Single and Multiple Threshold Models," *Journal of Econometrics*, 110, 319-352.

Gonzalo, J., and Pitarakis, J. (2012), "Regime Specific Predictability in Predictive Regressions," *Journal of Business and Economic Statistics*, 30, 229-241.

Goyal, A., and Welch, I. (2014), "Update to *A Comprehensive Look at the Empirical Performance of Equity Premium Prediction*," Unpublished Manuscript.

Hansen, B. E. (2000), "Sample splitting and threshold estimation," *Econometrica*, 68, 575-603.

Henkel, S. J., Martin, J. S., and Nardari, F. (2011), "Time-Varying Short-Horizon Predictability," *Journal of Financial Economics*, 99, 560-580.

Jansson, M., and Moreira, M. J. (2006), "Optimal Inference in Regression Models with Nearly Integrated Regressors," *Econometrica*, 74, 681-714.

Kim, J., and Pollard,D. (1990), "Cube Root Asymptotics," *Annals of Statistics*, 18, 191-219.

Kostakis, A., Magdalinos, A., and Stamatogiannis, M. (2014), "Robust Econometric Inference for Stock Return Predictability," *Review of Financial Studies,* forthcoming.

Lewellen, J. (2004), "Predicting returns with financial rations," *Journal of Financial Economics*, 74, 209-235.

Paye, B. S. and Timmermann, A. (2006), "Instability of Return Prediction Models," *Journal of Empirical Finance*, 13, 274-315.

Pesaran, M. H., and Timmermann, A. (1995), "Predictability of Stock Returns: Robustness and Economic Significance," *Journal of Finance,* 50, 1201-2228.

Phillips, P. C. B. (1988), "Regression Theory for Near-Integrated Time Series," *Econometrica*, 56, 1021-1043.

Phillips, P. C. B., and Magdalinos, A. (2009), "Econometric Inference in the Vicinity of Unity," *Singapore Management University, CoFie Working Paper No. 7.*

Pitarakis, J. (2008), "Threshold Autoregressions with a unit root: Comment," *Econometrica*, 76, 1207-1217.

Rossi, B., (2005), "Optimal Tests for Nested Model Selection with Underlying Parameter Stability," *Econometric Theory*, 21, 962-990.

Rossi, B., (2007), "Expectations Hypotheses Tests at Long Horizons," *Econometrics Journal*, 10, 1-26.

Valkanov, R. (2003), "Long-horizon regressions: theoretical results and applications," *Journal of Financial Economics,* 68, 201-232.

Wei, M., and Wright, J. (2013), "Reverse Regressions and Long Horizon Forecasting," *Journal of Applied Econometrics,* 28, 353-371.

Welch, I., and Goyal, A. (2008), "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21, 1455-1508.

White, H. (2000), *Asymptotic Theory for Econometricians.* Second Revised Edition. Academic Press.