Predictive Regressions[*]

Jesús Gonzalo
Universidad Carlos III de Madrid
Department of Economics
jesus.gonzalo@uc3m.es

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

November 20, 2018

[*] Prepared for the Oxford Research Encyclopedia in Economics and Finance.

1 Introduction

Predictive regressions refer to linear regression models designed to assess the predictive power of past values of some economic or financial variable for the future values of another variable. In their simplest univariate form these predictive regressions are formulated as

$$y_t \quad = \quad \beta_0 + \beta_1 \; x_{t-1} + u_t \tag{1}$$

with the main concern being the testing of the statistical significance of an estimate of β_1. Although such models are no different from standard simple linear regression specifications with lagged explanatory variables and for which traditional inferences should carry through under mild assumptions the specific context in which they are encountered in many Economics and Finance applications and the dynamic properties of commonly considered predictors in particular has led to a vast body of research aiming to improve the quality and accuracy of inferences in such settings. Indeed, across many applications involving the estimation of such predictive regressions it is often the case that predictors are highly persistent behaving like nearly non-stationary processes while predictands are typically noisier with rapidly mean-reverting dynamics instead. This imbalance in the stochastic properties of predictors and predictand is also often combined with the presence of sizeable contemporaneous correlations between the shocks driving y_t and x_t. The co-existence of these two important features and their common presence in many Economic and Finance applications tends to seriously distort inferences based on traditional significance tests that rely on standard normal approximations (e.g. t-ratios used on least squares based estimates of β_1).

One of the most commonly encountered empirical application that is subject to the above complications has originated in the asset pricing literature and has involved the study of the predictability of stock returns with valuation ratios and dividend yields in particular. Such predictors are well known to have roots close to unity in their autoregressive representation making shocks to such series last for very long periods instead of dying off quickly, hence their labelling as highly persistent predictors. Stock returns on the other hand are well known to have very short memory with virtually no serial correlation resulting in much noisier dynamics relative to those predictors. In parallel to these distinct stochastic characteristics of predictand and predictors it is also often the case that shocks to scaled price variables (e.g. price to earnings, price to book value, price to sales) are contemporaneously negatively correlated with shocks to returns.

The distortions that characterise traditional least squares and t-ratio based inferences conducted on β_1 in such settings typically materialise in the form of important size distortions that lead to too frequent wrong rejections of the null hypothesis and the finding of spurious predictability. It is important to emphasise however that these distortions are driven by the *joint* presence of persistence *and* contemporaneous correlations with the latter's magnitudes driving the seriousness of these wrong rejections.

These econometric complications have led to a vast research agenda aiming to develop alternative solutions to conducting inferences about β_1 with good size and power properties even under persistence and sizeable contemporaneous correlations between predictors and predictand. A very prolific avenue of research in this context has involved recognising the persistent nature of predictors by explicitly modelling them as nearly non-stationary local to unit root processes. Given that commonly used predictors such as valuation ratios cannot logically be viewed as pure unit root processes as this would imply that prices and fundamentals (e.g. earnings, dividends etc) can diverge for long periods the use of a near non-stationary framework offers a particularly useful compromise for capturing the stylised facts associated with these regression models. A popular specification for capturing persistence is the well known local to unit-root model often specified as

$$x_t = \left(1 - \frac{c}{T}\right) x_{t-1} + v_t \tag{2}$$

with c referring to a strictly positive constant and T to the sample size so that the associated autocorrelation coefficient is less than but possibly very close to unity. Such parameterisations lead to non-standard and non Gaussian asymptotics for the associated test statistics used to test hypotheses on β_1 and their implementation requires the use of simulation based critical values. These non-standard asymptotics can also easily accommodate contemporaneous correlations between u_t and v_t and it is generally hoped that they may lead to statistics with better size and power properties compared to the use of standard inferences relying on normal approximations.

One fundamental drawback of this more realistic framework however has to do with the fact that these non-standard asymptotics taking the form of stochastic integrals in Gaussian processes also depend on the unknown and non-estimable non-centrality parameter c used to model persistence thus making their practical implementation difficult. An important ensuing agenda then aimed at addressing this problem through more or less successful means. Early approaches involved considering bounds type tests that rely on multiple tests conducted over a range of value of the nuisance parameter and subsequently corrected using Bonferroni bounds (Cavanagh, Elliott and Stock (1995), Campbell and Yogo (2006), Jansson and Moreira (2006)). More recently the focus has shifted towards methods that involve either model or test statistic transformations so as to robustify the asymptotics to the influence of c. Examples include the use of instrumental variable as opposed to least squares based estimation of β_1 with instruments designed in such a way that the resulting asymptotics no longer depend on c (Phillips and Magdalinos (2009), Kostakis, Magdalinos and Stamatogiannis (2015)). Other related approaches have relied on model augmentation techniques that augment the original predictive regression with an additional predictor selected in such a way that inferences about β_1 have convenient nuisance parameter free distributions (Breitung and Demetrescu (2014)). These two approaches have now become the norm in the applied literature due to their good size and power properties and their ability to accommodate a rich set of features such as heteroskedasticity and serial correlation. A particularly useful feature of these methods is also their ability to handle multiple persistent predictors within (1) and to effectively be immune to persistence.

This line of research aiming at improving and robustifying inferences in the context of these predictive regressions also opened the way to novel approaches to modelling predictability and to the introduction of non-linearities in particular. The main motivation driving this important extension and generalisation was the recognition that predictability may not be a *stable* phenomenon but possibly varying across time or across economically relevant episodes. The predictive power of a predictor may for instance kick in solely during particular economic times while shutting off in other times (e.g. recessions versus expandions versus normal times). If ignored, the presence of such phenomena will almost certainly distort inferences about predictability in the sense of leading to conflicting outcomes depending on the sample periods being considered.

A burgeoning research agenda in this area has involved introducing the presence of regime specific non-linearities (e.g. structural breaks, threshold effects) within these predictive regressions while at the same time continuing to address the complications arising from the persistent nature of predictors and the particular type of endogeneity induced by the strong contemporaneous correlation between u_t and v_t. An early example of a nonlinear predictive regression model in which nonlinearities have been modelled via threshold effects has for instance been introduced in Gonzalo and Pitarakis (2012, 2017). This new class of threshold predictive regressions allowed the parameters of the model to potentially alternate between two possible values depending on whether a variable proxying for the economic cycle exceeds or lies below a threshold parameter. This offered a convenient and intuitive way of *attaching a cause* to the presence of predictability while also allowing it to shut off during particular periods. Another related extension has involved allowing the parameters of (1) to be subject to structural breaks with time effectively acting as a threshold variable. Pitarakis (2017) has introduced a battery of tests designed to detect the presence of such effects while at the same time addressing the two common econometric complications. A related modelling framework has also been recently developed in Farmer, Schmidt and Timmermann (2018) where the authors introduced the notion of *pockets of predictability* captured via smoothly varying functional parameters viewed as functions of time. Other fully non-parametric approaches effectively remaining agnostic about the functional firm linking y_t and x_{t-1} have also been developed in Juhl (2014), Kasparis, Andreou and Phillips (2015) amongst others.

2 Simple Predictive Regressions: Inference problems and Early Research

Operating within the simple specification given by (1)-(2) it is initially instructive to illustrate in greater depth the econometric complications that arise when testing the null hypothesis $H_0 : \beta = 0$ under the explicit modelling of the predictor as a near unit root process. For the sake of the exposition it is assumed that u_t and v_t are stationary disturbances that are i.i.d. but correlated and with the associated variance-covariance matrix given by $\Sigma = \{\{\sigma_u^2, \sigma_{uv}\}, \{\sigma_{uv}, \sigma_v^2\}\}$.

Given this simplified framework and some further regularity conditions (see Phillips (1987)) it is well known that the stochastic process $X_T(r) = x_{[Tr]}/\sqrt{T}$ where $x_{[Tr]} = \sum_{i=1}^{[Tr]}(1 - c/T)^{[Tr]-i}v_i$ satisfies an invariance principle with $X_T(r) \Rightarrow J_c(r)$ for $r \in [0, 1]$. Here $J_c(r)$ is referred to as an Ornstein-Uhlenbeck process and can informally be viewed as the continuous time equivalent of an autoregressive process. More specifically $J_c(r) = \int_0^r e^{(r-s)c}dW_v(s)$ with $W_v(r)$ denoting a standard Brownian Motion associated with the $v_t's$. This process is clearly Gaussian but with the complication that its variance depends on a DGP specific parameter, namely c. As an FCLT also holds for $w_t = (u_t, v_t)'$ with $T^{-\frac{1}{2}}\sum_{t=1}^{[Tr]} w_t \Rightarrow \Sigma^{\frac{1}{2}}(W_u(r), W_v(r))'$, following Cavanagh, Elliott and Stock (1995) the t-ratio associated with β_1 follows

$$t_{\hat{\beta}_1} \quad \Rightarrow \quad \rho\, \frac{\int_0^1 J_c(r)dW_v(r)}{\sqrt{\int_0^1 J_c(r)^2 dr}} + \sqrt{1-\rho^2}Z \tag{3}$$

where $\rho = \sigma_{uv}/\sigma_u\sigma_v$ and Z denotes a standard normal random variable.

The formulation in (3) is particularly instructive for understanding the nature of the complications that arise in predictive regressions and the joint role played by the presence of high persistence and a non-zero ρ (induced by the nonzero contemporaneous covariance σ_{uv}) in particular. In such instances the limiting distribution in (3) depends on the non-centrality parameter c complicating the practical implementation of inferences based on $t_{\hat{\beta}_1}$. If $\rho = 0$ however we have $t_{\hat{\beta}_1} \Rightarrow N(0, 1)$ suggesting that the normal approximation should lead to a test that is properly sized under sufficiently large sample sizes.

Early research in this area has addressed the problem of the dependence of inferences on c through a variety of methods which although theoretically sound were subject to practical shortcomings often leading to tests that were conservative and having low power. Given the dependence of the quantiles of the limiting distribution in (3) on the unknown noncentrality parameter c popular approaches relied on the early literature on multiple testing and Bonferroni based techniques in particular.

In Cavanagh, Elliott and Stock (1995) for instance the authors developed a Bonferroni based confidence interval for β_1 that relied on an intial confidence interval for c obtained following the confidence belt methodology of Stock (1991). Stock (1991)'s approach for constructing a confidence interval for c (equivalently $\phi = 1 - c/T$) involves first implementing an Augmented Dickey Fuller type t-test for testing $H_0 : \phi = 1$ on x_t. This ADF t-test is distributed as

$$\hat{t}_{adf}(c) \quad \Rightarrow \quad \frac{\int J_c(r) dW_v(r)}{\sqrt{\int J_c(r)^2 dr}} + \frac{c}{\sqrt{\int J_c(r)^2 dr}} \equiv t_{adf}(c) \tag{4}$$

which depends solely on c. The idea is then to use the duality between hypothesis testing and confidence intervals to obtain a confidence interval for c via the inversion of the acceptance region of the test. Letting $h_{L,\frac{\alpha_1}{2}}$ and $h_{U,1-\frac{\alpha_1}{2}}$ denote the $\alpha_1/2$ and $1 - \alpha_1/2$ percentiles of $t_{adf}(c)$ we can write $\hat{t}_{adf}(c) \in [h_{L,\frac{\alpha_1}{2}}, h_{U,1-\frac{\alpha_1}{2}}]$ for the acceptance region of the test statistic. These critical values can then be inverted numerically to lead to the confidence interval for c say $CI_c(\alpha_1) = [h^{-1}_{U,\frac{\alpha_1}{2}}(t_{adf}(c)), h^{-1}_{L,1-\frac{\alpha_1}{2}}(t_{adf}(c))] \equiv [c_L(\alpha_1), c_U(\alpha_1)]$ which is obtained for some given value of the test statistic and which effectively provides the range of values of c that are in the above acceptance region. For each value of c in this interval one can subsequently construct confidence intervals for β_1 using the limiting distribution of $t_{\hat{\beta}_1}$ in (3). It is worth pointing out however that these confidence intervals for c have some undesirable properties in the sense of not being uniform in ϕ and leading to generally poor outcomes when the underlying ϕ is too far off the unit root scenario (see Mikusheva (2007) who proposed an alternative way of constructing these confidence intervals for c using a modification to the $\hat{t}_{adf}(c)$ statistic that leads to confidence intervals that are uniform across ϕ). Given the confidence interval for c it is then possible to proceed with a Bonferroni based approach to obtain a confidence interval for β_1 that no longer depends on c. More formally a confidence interval for β_1 is first constructed for each value of c, say $CI_{\beta_1|c}(\alpha_2)$ using the limiting distribution of $t_{\hat{\beta}_1}$ in (3). A final confidence interval for β_1 that does not depend on c is then obtained as the union across $c \in [c_L(\alpha_1), c_U(\alpha_2)]$ of these $CI_{\beta_1|c}(\alpha_2)'$s leading to $CI_{\beta_1}(\alpha_1, \alpha_2) = [\min_{c_L(\alpha_1) \leq c \leq c_U(\alpha_1)} d_{t_{\hat{\beta}_1},c,\frac{\alpha_2}{2}}, \max_{c_L(\alpha_1) \leq c \leq c_U(\alpha_1)} d_{t_{\hat{\beta}_1},c,\frac{1-\alpha_2}{2}}]$ with $d_{t_{\hat{\beta}_1},c}$ referring to the critical values associated with (3).

Within the above methodological context it is important to recognise that the choice of using the ADF based t-ratio for obtaining a confidence interval for c followed by the use of $t_{\hat{\beta}_1}$ is arbitrary in the sense that alternative test statistics fulfilling the same purpose may also be considered. An important literature followed this line of research by considering alternative test statistics with better optimality properties and better power properties across the relevant range of ϕ (see Elliott and Stock (2001) for instance for an alternative approach to obtaining confidence intervals for c that relies on the the point optimal test proposed in Elliott, Rothemberg and Stock (1996)). In an influential paper Campbell and Yogo (2006) focused on these issues in the specific context of the predictive regression setting as in (1)-(2). For the construction of a confidence interval for c they proposed to rely on the more efficient DF-GLS test of Elliott, Rothemberg and Stock (1996) and for which they provided tabulations linking the magnitude of this test statistic with a corresponding confidence interval for c. Given this alternative approach to obtaining the relevant range of c values they subsequently also introduced an alternative to $t_{\hat{\beta}_1}$ which they referred to as their Q statistic. The latter is effectively a t-ratio on β_1 but obtained from the augmented specification $y_t = \beta_1 x_{t-1} + \lambda(x_t - \phi x_{t-1}) + \eta_t$ with $\lambda = \sigma_{uv}/\sigma_v^2$ and shown to lead to better power properties compared to the use of $t_{\hat{\beta}_1}$.

An important limitation of all these *two stage* confidence interval based approaches is that the resulting confidence intervals are typically not uniform in ϕ, have potentially zero coverage probabilities and may lead to subsequently poor power properties when it comes to conducting inferences about β_1. An excellent technical discussion of these shortcomings can be found in Phillips (2015). Also noteworthy is the fact that these methods are difficult to generalise to multi-predictor settings or for handling more flexible assumption on the variances of the errror processes.

Alternative routes to improving inferences about β_1 within (1)-(2) have also been considered around the same time as the above early literature. One line of research involved improving the quality of the least squares estimator of β_1 by removing its bias. Note for instance that the least squares estimator of β_1 obtained from (1) is not unbiased as the predictor is not strictly exogenous. As shown in Stambaugh (1999) the bias of $\hat{\beta}_1$ can be formulated as $E[\hat{\beta}_1 - \beta_1] = (\sigma_{uv}/\sigma_v^2)E[\hat{\phi} - \phi]$. Under $\phi \approx 1$ it is well known that $\hat{\phi}$ has a strong downward bias and as σ_{uv} is typically negative one clearly expects a strong upward bias in $\hat{\beta}_1$. It is this undesirable feature of $\hat{\beta}_1$ that this literature has attempted to address by appealing to existing results on biases of first order autocorrelation coefficients (e.g. Kendall (1954)) such as $E[\hat{\phi}-\phi] = -(1+3\phi)/T+O(T^{-2})$. We note for instance that an adjusted estimator of the slope parameter $\hat{\beta}_1^c = \hat{\beta}_1 + (\sigma_{uv}/\sigma_v^2)((1+3\phi)/T)$ satisfies $E[\hat{\beta}_1^c - \beta_1] = 0$ under known ϕ. Lewellen (2003) took advantage of these results to devise an alternative approach to testing $H_0 : \beta_1 = 0$ that relies on a bias corrected estimator of β_1 given by $\hat{\beta}_1^{lw} = \hat{\beta}_1 + (\sigma_{uv}/\sigma_v^2)(\hat{\phi} - \phi)$ with ϕ set at 0.9999. This expression is subsequently operationalised by replacing σ_{uv} and σ_v^2 by suitable estimates. Naturally these approaches rely on an important set of assumptions for their validity (e.g. normality) and require a certain level of ad-hoc input.

3 Robustifying Inferences to the noncentrality parameter: Recent Developments

A more recent trend in this literature on conducting inferences in predictive regressions with persistent predictors has aimed to jointly address two key concerns. The first concern is the need to operate within a more flexible environment than (1)-(2) that can accommodate multiple predictors while also taking into account complications such as serial correlation and heteroskedasticity. The second concern stems from the need to develop inferences with good size and power properties that are also robust to the persistence properties of the predictors. A more empirically relevant generalisation of (1)-(2) can for instance be formulated as

$$
\begin{aligned}
y_t &= \boldsymbol{\beta}'\boldsymbol{x}_{t-1} + u_t && (5)\\
\boldsymbol{x}_t &= (I_p - \boldsymbol{C}/T)\boldsymbol{x}_{t-1} + v_t && (6)
\end{aligned}
$$

with $\boldsymbol{C} = diag(c_1, \ldots, c_p)$, $c_i > 0$ and u_t and v_t modelled as possibly dependent and cross-correlated stationary processes.

A novel approach to the problem of estimating the parameters of (5) and testing relevant hypotheses on $\boldsymbol{\beta}$ has beeen developed in Kostakis, Magdalinos and Stamatogiannis (2015) where the authors introduced an instrumental variable based approach designed in such a way that the resulting asymptotics of a suitably normalised Wald statistic for testing hypotheses of the form $H_0 : \boldsymbol{R\beta} = r$ in (5) are standard χ^2 and not dependent on the $c_i's$. Their framework is in fact more general than (5)-(6) as it can also accommodate predictors that are more or less persistent than those modelled as in (6) including pure unit-root, stationary or mildly persistent processes parameterised as $x_{it} = (1 - c_i/T^{\alpha})x_{it-1} + v_{it}$ with $\alpha \in (0, 1)$. The strength of the methodology lies in the fact that one can effectively operate and conduct inferences about $\boldsymbol{\beta}$ while being agnostic about the degree of persistence of the predictors. Its reliance on a standard Wald statistic also makes the implementation of traditional Newey-West type corrections particularly straightforward. This instrumental variable approach has originated in the earlier work of Phillips and Magdalinos (2009) who focused on a multivariate cointegrated system closely related to (5)-(6) with (5) replaced with $y_t = \boldsymbol{\beta}'\boldsymbol{x}_t + u_t$ and labelled as a cointegrated system with persistent predictors.

The main idea behind the instrumental variable approach involves instrumenting $\boldsymbol{x}_t$ with a slightly less persistent version of itself constructed with the help of the first differenced $\boldsymbol{x}_t's$. In this sense the IV is generated using solely model sepecific information and does not require any external information, hence its labelling as IVX. More specifically the p-vector of instruments for

$\boldsymbol{x}_t$ is constructed as

$$\widetilde{\boldsymbol{z}}_t \;=\; \sum_{j=1}^{t}\left(I_p - \boldsymbol{C}_z/T^{\delta}\right)^{t-j}\Delta\boldsymbol{x}_j \tag{7}$$

for a given $\delta \in (0,1)$ and some given $\boldsymbol{C}_z = diag(c_{z1},\dots,c_{zp})$, $c_{z,i} > 0$ for $i = 1,\dots,p$. Note that as $\delta < 1$ the instruments are less persistent than $\boldsymbol{x}_t$. From (6) we have $\Delta\boldsymbol{x}_j = -\boldsymbol{C}/T + v_t$ which when combined within (7) leads to the following decomposition of the instrument vector

$$\widetilde{\boldsymbol{z}}_t \;=\; \boldsymbol{z}_t - \frac{\boldsymbol{C}}{T}\Psi_t \tag{8}$$

with $\boldsymbol{z}_t = \sum_{j=1}^{t}(I_p - \boldsymbol{C}_z/T)^{t-j}v_j$ and $\Psi_t = \sum_{j=1}^{t}(I_p - \boldsymbol{C}/T)^{t-j}\boldsymbol{x}_{j-1}$. Note that $\boldsymbol{z}_t$ is such that $\boldsymbol{z}_t = (I_p - \boldsymbol{C}_z/T^{\delta})\boldsymbol{z}_{t-1} + v_t$ while Ψ_t is a remainder term shown not to have any influence on the asymptotics. For practical purposes these mildly integrated IVs are generated as a filtered version of $\boldsymbol{x}_t$ using (7) with a given $\boldsymbol{C}_z$ and δ and are approximately equivalent to $\boldsymbol{z}_t$. These are then used to obtain an IV based estimator of $\boldsymbol{\beta}$ from (5). More formally, letting $\boldsymbol{X}$ and $\boldsymbol{Z}$ denote the regressor and IV matrices respectively, both obtained by stacking the elements of $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ we have $\hat{\boldsymbol{\beta}}^{ivx} = (\boldsymbol{X'Z})^{-1}\boldsymbol{Z'}y$ and the associated conditionally homoskedastic version of the Wald statistic is given by

$$W_T \;=\; (\boldsymbol{R}\hat{\boldsymbol{\beta}}^{ivx} - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{Z'X})^{-1}(\boldsymbol{Z'Z})(\boldsymbol{X'Z})^{-1}\boldsymbol{R'}]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}^{ivx} - \boldsymbol{r})/\widetilde{\sigma}_u^2 \tag{9}$$

with $\widetilde{\sigma}_u^2 = \sum_t(y_t - \hat{\boldsymbol{\beta}}^{ivx\prime}\boldsymbol{x}_t)^2/T$. In the context of (5)-(6) Kostakis, Magdalinos and Stamatogiannis (2015) established that

$$W_T \;\Rightarrow\; \chi^2(m) \tag{10}$$

with m referring to the rank of the restriction matrix $\boldsymbol{R}$, thus removing the need to be concerned with the magnitude of the $c_i's$ that parameterise the persistent predictors in the DGP.

Here it is important to point out that $\hat{\boldsymbol{\beta}}^{ivx}$ continues to have a limiting distribution that depends on the c_i's so that the strength of the IVX methodology operates via the Wald statistic's variance normalisation as illustrated by the middle term in (9) and which effectively *cancels out* the asymptotic variance of $\hat{\boldsymbol{\beta}}^{ivx}$ leading to an identity matrix (due to the asymptoic mixed normality of $\hat{\boldsymbol{\beta}}^{ivx}$). Note also that the use of this IV approach is not inconsequential for the asymptotic properties of $\hat{\boldsymbol{\beta}}^{ivx}$ which converges at a rate slower than $\hat{\boldsymbol{\beta}}$ with a rate determined by the magnitude of δ used in the construction of the IVs. More specifically $\hat{\boldsymbol{\beta}}^{ivx} - \boldsymbol{\beta} = O_p(T^{-\frac{1+\delta}{2}})$ which can be compared with the T-consistency of the standard least squares estimator $\hat{\boldsymbol{\beta}}$.

To highlight some of these properties more explicitly it is useful to revisit the simple univariate setting of (1)-(2). Letting y_t^*, x_t^* and $\widetilde{z}_t^*$ denote the demeaned versions of the variables of interest the IVX based estimator of β_1 is given by

$$\hat{\beta}_1^{ivx} \quad = \quad \frac{\sum_t y_t^* \widetilde{z}_{t-1}^*}{\sum_t x_{t-1}^* \widetilde{z}_{t-1}^*} \tag{11}$$

and from Phillips and Magdalinos (2009) and Kostakis, Magdalinos and Stamatogiannis (2015) we have

$$T^{\frac{1+\delta}{2}}(\hat{\beta}_1^{ivx} - \beta_1) \quad \Rightarrow MN\left(0, \frac{c_z \sigma_u^2 \sigma_v^2}{2(\sigma_v^2 + \int_0^1 J_c(r) dJ_c(r))}\right) \tag{12}$$

which highlights the fact that the distribution of the IVX estimator continues to depend on c via the presence of the $J_c(r)$ process in the asymptotic variance. Thanks to the mixed normality of $\hat{\beta}_1^{ivx}$ however the use of the IV based variance normalisation embedded in the Wald statistic given here by

$$W_T(\beta_1 = 0) \quad = \quad \frac{(\hat{\beta}_1^{ivx})^2 \sum x_{t-1}^* \widetilde{z}_{t-1}^*}{\widetilde{\sigma}_u^2 \sum (\widetilde{z}_{t-1}^*)^2} \tag{13}$$

leads to the outcome that $W_T(\beta_1 = 0) \Rightarrow \chi^2(1)$. At this stage it is also useful to point out that the demeaning of the IVs used in (11) and (13) was not strictly necessary as the IVX based estimator of β_1 is invariant to their demeaning as discussed in Kostakis, Magdalinos and Stamatogiannis (2015)

The implementation of the $\hat{\beta}_1^{ivx}$ estimator requires one to take a stance on the magnitudes of c_z and δ which are needed for generating the instrumental variables. As the choice of c_z is innocuous Phillips and Magdalinos (2009) suggest setting $c_z = 1$. The impact of δ used in the construction of the IVs is more problematic however. Although the asymptotic analysis requires $\delta \in (0, 1)$ it is clear that a choice for δ that is close to 1 will make the IV closer to the original variable it is instrumenting while choosing a δ much lower than 1 will have the opposite effect. As the choice of δ directly influences the rate of convergence of $\hat{\boldsymbol{\beta}}^{ivx}$ with lower magnitudes of δ implying a slower rate of convergence it is natural to expect that the choice of δ may raise important size versus power tradeoffs, in smaller samples in particular. Kostakis, Magdalinos and Stamatogiannis (2015) argue that a choice such as $\delta = 0.95$ offers excellent size/power tradeoffs while they advise against choosing $\delta < 0.9$ due to potentially negative power implications. As shown in their simulations the closer δ is to 1 the better the power properties of the IVX based Wald statistic. However this choice also tends to create non ignorable size distortions in moderate sample sizes such as $T = 500$. This is an issue the authors have explored in great detail showing that the estimation of an intercept in (5) is the key driver of these size distortions that further amplify as $\delta \to 1$. To remedy this problem they introduced a finite sample correction to the formulation of the Wald statistic in (9) and that is shown to make the Wald statistic match its asymptotic limit very accurately in finite samples even for δ close to 1. Note also that the inclusion of this finite sample correction has no bearing on the χ^2 asymptotics in (10). In the context of the formulation in (9) the middle term $\widetilde{\sigma}^2_u \boldsymbol{Z}'\boldsymbol{Z}$ of the quadratic form is replaced with $\widetilde{\sigma}^2_u[\boldsymbol{Z}'\boldsymbol{Z} - \overline{\boldsymbol{z}}_T \ \overline{\boldsymbol{z}}'_T(1-\hat{\gamma})]$ with $\hat{\gamma} = \hat{\sigma}^2_{uv}/\hat{\sigma}^2_u\hat{\sigma}^2_v$ and $\overline{\boldsymbol{z}}_T$ referring to the p-vector of sample means of the IVs. This simple correction is shown to lead to a Wald statistic with excellent size control and power across a very broad range of persistent parameters.

An alternative yet similar approach to handling inferences within models such as (5)-(6) was also introduced in Breitung and Demetrescu (2015) who focused on a model augmentation approach instead. In the context of a simple predictive regression the idea behind variable augmentation is to expand the specification in (1) with an additional carefully chosen regressor and testing $H_0 : \beta_1 = 0$ in

$$y_t \quad = \quad \beta_1 \ z_{t-1} + \psi_1 \ (x_{t-1} - z_{t-1}) + u_t \tag{14}$$

ignoring the restriction $\beta_1 = \psi_1$. They subsequently show that choosing z_t to satisfy a range of characteristics including that it is less persistent than x_t leads to standard normally distributed t-ratios despite the presence of the highly persistent predictor in the DGP. These characteristics effectively require z_t and related cross-moments to satisfy law of large numbers and CLT type results (e.g. for $\eta \in [0, 1/2]$, $\sum z^2_{t-1}/T^{1+2\eta} = O_p(1)$, $\sum z^2_{t-1}u^2_t/T^{1+2\eta} \Rightarrow V_{zu} = O_p(1)$, $\sum z_{t-1}x_{t-1}/T^{\frac{3}{2}+\eta} \xrightarrow{p} 0$ and $\sum z_{t-1}u_t/\sqrt{V_{zu}}T^{\frac{1}{2}+\eta} \Rightarrow N(0,1)$). Inferences can be conducted using a t-statistic that can be further corrected for heteroskedasticity à la Eicker-White. There is naturally a broad range of candidates for z_t that satisfy the above requirements including for instance the IVX variable of Phillips and Magdalinos (2009) but also fractionally integrated processes, short memory processes etc. As discussed by the authors these choices may have important implications for the power properties of the tests. The framework in (14) can also be straightforwardly adapted to include both deterministic components such as an intercept and trends and multiple predictors as in (5) leading to χ^2 distributed Wald statistics for testing $H_0 : \boldsymbol{\beta} = 0$.

4 Capturing Non-linearities within Predictive Regressions

This vast body of research on predictive regressions has mainly operated within a linear setting implying that predictability if present is a *stable* phenomenon in the sense that the full sample based estimator of $\hat{\boldsymbol{\beta}}$ converges to its true and potentially non-zero counterpart $\boldsymbol{\beta}$. This naturally rules out scenarios whereby predictability may be a time varying phenomenon with periods during which $\boldsymbol{\beta} = 0$ and periods during which $\boldsymbol{\beta} \neq 0$. Ignoring such economically meaningful phenomena may seriously distort the validity of the above techniques and the reliability of conclusions about the presence or absence of predictability. In the context of the predictability of stock returns for instance the presence of such phenomena may explain the conflicting empirical results that have appeared in the applied literature depending on the sample periods being considered.

These concerns have led to a novel research agenda that aimed to explicitly account for potential time variation in predictability by considering predictive regressions specified as

$$y_t \quad = \quad \boldsymbol{\beta}_t{}' \boldsymbol{x}_{t-1} + u_t \tag{15}$$

with $\boldsymbol{x}_t$ as in (6). Naturally this more realistic and flexible setting raises its own difficulties as one needs to take a stance on the type of time variation driving the evolving parameters. Popular parametric approaches that have been considered in the literature include standard structural breaks, threshold effects amongst others. All of these regime specific approaches effectively model time variation as

$$\boldsymbol{\beta}_t \quad = \quad \boldsymbol{\beta}_1 * D_t + \boldsymbol{\beta}_2 * (1 - D_t) \tag{16}$$

with D_t referring to a suitable 0/1 dummy variable. Such specifications allow predictability to shut-off over particular periods (e.g. $\boldsymbol{\beta}_1 = 0$ and $\boldsymbol{\beta}_2 \neq 0$) determined by the way the dummy variables have been defined making hypotheses such as $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ or $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = 0$ important to assess and provide a toolkit for.

Most of this literature has operated within simple univariate settings with only limited results developed for the multi-predictor case. In Gonzalo and Pitarakis (2012, 2017) the authors argued that a threshold based parameterisation of (16) can provide an economically meaningful yet parsimonious way of modelling time variation in the $\boldsymbol{\beta}$'s. The inclusion of threshold effects effectively turns the linear predictive regressions into a piecewise linear processes in which regimes are determined by the magnitude of a suitable threshold variable selected by the investigator. More formally within the simple predictive regression context a two-regime threshold specification comforming to the notation in (16) can be formulated as

$$y_t \quad = \quad (\beta_{01} I(q_{t-1} \leq \gamma) + \beta_{02} I(q_{t-1} > \gamma)) + (\beta_{11} I(q_{t-1} \leq \gamma) + \beta_{12} I(q_{t-1} > \gamma)) x_{t-1} + u_t \tag{17}$$

where q_t is an observed threshold variable whose magnitude relative to γ determines the regime structure. If q_t is taken as a proxy of the business cycle for instance the above specification could allow predictability to kick in (or be weaker/stronger) across economic episodes such as expansions and recessions. The fact that the threshold variable q_t is under the *control* of the investigator can also be viewed as particularly advantageous in this context as it allows one to attach an observable cause to what drives time variation in predictability. An important additional advantage of using piecewise linear structures such as (17) comes from the fact that such functions may provide good approximations for a much wider class of functional forms as demonstrated in Petruccelli (1992).

In Gonzalo and Pitarakis (2012) the authors focused on predictive regressions of the type presented in (17) with their stochastic properties assumed to mimic the environments considered in the linear predictive regression literature (e.g. allowing for persistence and endogeneity). The threshold variable q_t was in turn modelled as a strictly stationary and ergodic process whose innovations could potentially be correlated with those driving the predictor and predictand. Despite the presence of a highly persistent predictor parameterised as a local to unit root process Gonzalo and Pitarakis (2012) showed that a Wald type statistic for testing the null hypothesis of linearity ($H_0 : (\beta_{01}, \beta_{11}) = (\beta_{02}, \beta_{12})$) follows a well known distribution that is free of nuisance parameters and more importantly not dependent on c. As the framework in (17) also raises the issue of unidentified nuisance parameters (in this instance γ) under the null hypothesis inferences are conducted using supremum Wald type statistics viewed as a function of the unknown threshold parameter γ. Under suitable assumptions on the density of q_t the above indicator functions satisfy $I(q_t \leq \gamma) = I(F(q_t) \leq F(\gamma))$ with $F(.)$ denoting the distribution function of q_t so that the Wald statistic can also be viewed as a function of $F(\gamma) \equiv \lambda$ for purely technical reasons. The key result in Gonzalo and Pitarakis (2012) is given by

$$\sup_{\lambda \in (0,1)} W_T(\lambda) \Rightarrow \sup_{\lambda \in (0,1)} \frac{(B(\lambda) - \lambda B(1)'(B(\lambda) - \lambda B(1)}{\lambda(1-\lambda)} \tag{18}$$

with $B(\lambda)$ denoting a standard Brownian Motion whose dimension is given by the number of parameters whose equality is being tested under the null. A remarkable property of the limiting distribution in (18) is its robustness to the local to unit root parameter c, making inferences straightforward to implement. The task is further facilitated by the fact that the above limit can be recognised as a normalised vector Brownian Bridge and is extensively tabulated in the literature (see for instance Andrews (1993)). It is also important to note that the result in (18) remains valid in the context of (5)-(6) involving multiple predictors with potentially different $c_i^{\prime}s$. A rejection of the null hypothesis of linearity in (17) would clearly support the presence of regime specific predictability of y_t.

Another hypothesis of interest in this context is the joint null $H_0 : \beta_{01} = \beta_{02}, \beta_{11} = \beta_{12} = 0$ whose failure to be rejected would support a martingale difference type of behaviour for stock returns. Unlike the scenario in (18) however the Wald statistic associated with this latter hypothesis has a limiting distribution that depends on c and for which Gonzalo and Pitarakis (2012) developed an IVX type Wald statistic. More specifically, they showed that the Wald statistic for testing $H_0 : \beta_{01} = \beta_{02}, \beta_{11} = \beta_{12} = 0$ in (17) is asymptotically equivalent to the sum of two independent Wald statistics, with the first one given by $W_T(\lambda)$ in (18) used for testing $H_0 : (\beta_{01}, \beta_{11}) = (\beta_{02}, \beta_{12})$ and the second one associated with testing $H_0 : \beta_1 = 0$ in the linear predictive regression in (1) and for which an IVX procedure can be implemented, say $W_T^{ivx}(\beta_1 = 0)$ known to be distributed as $\chi^2(1)$. This allowed them to construct a novel statistic given by the sum of these two Wald statistics $\sup_\lambda W_T(\lambda) + W_T^{ivx}(\beta_1 = 0)$ and shown to be distributed as $\sup_\lambda (B(\lambda) - \lambda B(1))'(B(\lambda) - \lambda B(1))/\lambda(1-\lambda) + \chi^2(1)$. Although non-standard this limit is free of the influence of c and can be easily tabulated via simulation methods.

A rejection of these joint null hypotheses is naturally problematic to interpret when one is solely interested in whether regime specific predictability is induced by the highly persistent predictor x_t. This is because a rejection of the null may occur not because of shifting slope parameters but due to shifting intercepts instead (i.e. $\beta_{01} \neq \beta_{02}$). This issue has been subsequently addressed in Gonzalo and Pitarakis (2017) where the authors developed a Wald type test statistic for $H_0 : \beta_{11} = \beta_{12} = 0$ designed in such a way that its large sample behaviour remains robust to whether $\beta_{01} = \beta_{02}$ or $\beta_{01} \neq \beta_{02}$. Their method effectively relies on obtaining a conditional least squares based estimator of the unknown threshold parameter obtained from the null restricted version of (17) and using it as a plug-in estimator within an IVX based Wald statistic for testing $H_0 : \beta_{11} = \beta_{12} = 0$. This is then shown to be distributed as $\chi^2(2)$ under the null regardless of whether the threshold parameter estimator is spurious or consistent for an underlying true value i.e. regardless of whether the DGP has threshold effects in its intercept.

Other parametric alternatives to the threshold based approach have also been considered in this literature. A popular setting involves for instance allowing the parameters of the predictive regression to be subject to deterministic structural breaks, effectively replacing $I(q_t \leq \gamma)$ with $I(t \leq k)$ in (17). Due to the presence of the highly persistent predictor standard results from the structural break literature no longer apply in this context. Testing the null hypothesis of linearity via a SupWald type statistic for instance no longer follows the normalised Brownian Bridge type distribution tabulated in Andrews (1993). Unlike the simplifications that occur in the context of threshold effects and that lead to convenient outcomes as in (18) the main issue in this context continues to be the dependence of inferences on the unknown noncentrality parameter c with processes such as $J_c(r)$ appearing in the asymptotics. The invalidity of traditional parameter constancy tests under persistent predictors was pointed out in Rapach and Wohar (2006) who were concerned with assessing the presence of breaks in return based predictive regressions. In this early work they suggested using Hansen (2000)'s fixed regressor bootstrap as a way of controlling for the unknown degree of persistence in the predictors. This idea has also been taken up and expanded in the more recent work of Georgiev et al. (2018).

Pitarakis (2017) proposed to bypass some of these difficulties by developing a CUSUMSQ type statistic based on the squared residuals from (1) and shown to have a limiting distribution that does not depend on c as in

$$\max_{1\leq k\leq T} \frac{1}{\hat{\phi}_T}\left|\frac{\sum_{t=1}^{k}\hat{u}_t^2}{\sqrt{T}} - \frac{k}{T}\frac{\sum_{t=1}^{T}\hat{u}_t^2}{\sqrt{T}}\right| \quad \Rightarrow \quad \sup_{\pi\in[0,1]}|B(\pi)-\pi B(1)| \tag{19}$$

with $\hat{\phi}_T$ denoting a consistent estimator of the long run variance of $(u_t^2-\sigma_u^2)$. Here the $\hat{u}_t$'s refer to the standard least squares based residuals obtained from (1). The results obtained in Pitarakis (2017) naturally extend to multiple predictor settings (e.g. with $\hat{u}_t^2$ obtained from (5)), can accommodate conditional heteroskedasticity and have been shown to have excellent power properties with good size control. In related recent work Georgiev et al. (2018) also developed new inference methods within predictive regressions as in (5)-(6) with either stochastically (e.g. $\boldsymbol{\beta}_t$ evolving as a random walk) or deterministically varying (e.g. structural breaks) parameters using LM and SupWald type test statistics respectively. Their approach to neutralising the dependence of their asymptotics on the $c_i's$ relied on a fixed regressor bootstrapping algorithm that use the realised $x_{t-1}'s$ as a fixed regressor in the bootstrap.

The above parametric approaches for capturing nonlinearities have led to various novel stylised facts on the predictability of stock returns. Within the threshold setting of Gonzalo and Pitarakis (2012) for instance the authors documented strong countercyclicality in the predictability of US returns with dividend yields with the latter entering the predictive regression significantly solely during recessions. This phenomenon has generated considerable recent interest with numerous novel contributions aiming to explain it and document it more comprehensively. A particularly interesting novel approach has been introduced in Farmer, Schmidt and Timmermann (2018) in which the authors establish that *pockets of predictability* are a much broader phenomenon that is not solely confined to recessionary periods.

The concern for functional form mispecification that may affect the above parametric nonlinear settings has also motivated fully nonparametric approaches to assessing predictability by letting x_{t-1} enter (1) via an unknown functional form as in $y_t = f(x_{t-1}) + u_t$. A particularly useful approach has been developed in Andreou, Kasparis and Phillips (2015) where the authors focused on designing tests of $H_0 : f(x) = \mu$ based on the Nadaraya-Watson kernel regression estimator of $f(.)$ and whose distributions have been shown to be robust to the persistence properties of x_t including local to unit root parameterisations. One shortcoming of these nonparametric techniques is the weakness of their power properties against linear alternatives when compared with parametric approaches.

5 Further Remarks

This vast body of research broadly labelled as *predictive regression literature* has been driven by concerns that arose in empirical applications across a variety of fields and asset pricing in particular. Numerous new avenues of research that may help address novel questions through new methodological developments are expected to continue to further grow and enrich the existing literature. Given the increased availability of big data sets the issue of handling multiple predictors having different stochastic properties in either linear or nonlinear contexts will continue to create many technical challenges if one wishes to take advantage of the growing literature on high dimensional estimation, model selection and prediction via shrinkage based techniques. Most of this predictability literature has also been confined to the conditional mean of the predictands of interest whereas predictability may be a much broader phenonomenon potentially also (or solely) affecting quantiles of the series of interest. In numerous risk related applications one may for instance be interested in uncovering factors influencing the extreme tails of a series. Generalising the existing literature to accommodate time variation in such settings will almost certainly raise many novel challenges.

6 References

Andrews, D. W. K., (1993). Tests for Parameter Instability and Structural Change with Unknown Changepoint, *Econometrica*, Vol. 61, pp. 821-856.

Breitung, J., and Demetrescu, M., (2015). Instrumental variable and variable addition based inference in predictive regressions, *Journal of Econometrics*, 187, 358-375.

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

Cavanagh, C. L., Elliott, G., and Stock, J. H., (1995). Inference in Models with Nearly Integrated Regressors, *Econometric Theory*, 11, 1131-1147.

Elliott, G., Rothemberg, T. J., and Stock, J. H., (1996). Efficient Test for an Autoregressive Unit Root, *Econometrica*, 64, 813-836.

Elliott, G., and Stock, J. H., (2001). Confidence Intervals for Autoregressive Coefficients Near One, *Journal of Econometrics,* 103, 155-181.

Farmer, L., Schmidt, L., and Timmermann, A., (2018). Pockets of Predictability, SSRN Working Paper No. 3152386.

Georgiev, I., Harvey, D. I., Leybourne, S. J., Taylor, R. A. M., (2018). Testing for Parameter instability in predictive regression models, *Journal of Econometrics*, 204, 101-118.

Gonzalo, J., and Pitarakis, J., (2012). Regime specific predictability in predictive regressions, *Journal of Business and Economic Statistics*, 30, 229-241.

Gonzalo, J., and Pitarakis, J., (2017). Inferring the Predictability Induced by a Persistent Regressor in a Predictive Threshold Model,*Journal of Business and Economic Statistics*, 35, 202-217.

Hansen, B. E., (2000). Testing For Structural Change in Conditional Models. *Journal of Econometrics* 97, 93115.

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

Juhl, T., (2014). A Nonparametric Test of the Predictive Regression Model, *Journal of Business and Economic Statistics*, 32, 387-394.

Kasparis, I., Andreou, E., and Phillips, P.C.B, (2015). Nonparametric predictive regression, *Journal of Econometrics*, 185, 468-494.

Kendall, M.G., (1954). Note on bias in the estimation of autocorrelation. *Biometrika* 41, 403-404.

Kostakis, A., Magdalinos, A., and Stamatogiannis, M., (2015). Robust Econometric Inference for Stock Return Predictability, *Review of Financial Studies,* Vol. 28, pp. 1506-1553.

Lewellen, J., (2004). Predicting returns with financial ratios, *Journal of Financial Economics* 74, 209235.

Mikusheva, A., (2007). Uniform Inference in Autoregressive Models, *Econometrica*, 75, 1411-1452.

Petruccelli, J., (1992). On the approximation of time series by threshold autoregressive models, *Sankhya*, 54, 106-113.

Phillips, P. C. B., (1987). Time Series Regression with a Unit Root, *Econometrica*, Vol. 55, pp. 227-301.

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

Phillips, P. C. B., (2015). Pitfalls and possibilities in predictive regression, *Cowles Foundation Discussion Paper*, No. 2003.

Pitarakis, J., (2017). A Simple Approach for Diagnosing Instabilities in Predictive Regressions, *Oxford Bulletin of Economics and Statistics*, 79, 851-874.

Rapach, D. E., and Wohar, M. E., (2006). Structural Breaks and Predictive Regression Models of Aggregate U.S. Stock Returns, *Journal of Financial Econometrics*, 4, 238-274.

Stambaugh, R. F., (1999). Predictive Regressions, *Journal of Financial Economics*, 54, 375-421.

Stock, J. H., (1991). Confidence intervals for the largest autoregressive root in U.S. economic time series, *Journal of Monetary Economics*, 28, 435-460.

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