Cowles Foundation Discussion Papers
Cowles Foundation

6-1-2015

Pitfalls and Possibilities in Predictive Regression

Peter C.B. Phillips

Recommended Citation

Phillips, Peter C.B., "Pitfalls and Possibilities in Predictive Regression" (2015). *Cowles Foundation Discussion Papers*. 2438.
https://elischolar.library.yale.edu/cowles-discussion-paper-series/2438

PITFALLS AND POSSIBILITIES IN PREDICTIVE REGRESSION

By

Peter C. B. Phillips

June 2015

COWLES FOUNDATION DISCUSSION PAPER NO. 2003

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Pitfalls and Possibilities in Predictive Regression[*]

Peter C. B. Phillips
Yale University, University of Auckland,
Singapore Management University & University of Southampton

June 12, 2015

Abstract

Financial theory and econometric methodology both struggle in formulating models that are logically sound in reconciling short run martingale behaviour for financial assets with predictable long run behavior, leaving much of the research to be empirically driven. The present paper overviews recent contributions to this subject, focussing on the main pitfalls in conducting predictive regression and on some of the possibilities offered by modern econometric methods. The latter options include indirect inference and techniques of endogenous instrumentation that use convenient temporal transforms of persistent regressors. Some additional suggestions are made for bias elimination, quantile crossing amelioration, and control of predictive model misspecification.

Keywords: Bias, Endogenous instrumentation, Indirect inference, IVX estimation, Local unit roots, Mild integration, Prediction, Quantile crossing, Unit roots, Zero coverage probability.

JEL classification: C22, C23

"The records of 11 leading financial periodicals and services since 1927, over periods varying from 10 to 15$\frac{1}{2}$ years, fail to disclose evidence of ability to predict successfully the future course of the stock market." Alfred Cowles (1944).

---

[*] Presented as the Halbert White Memorial Lecture at the SoFiE conference held in Singapore at Singapore Management University, June 2013. The author acknowledges support from the NSF under Grant No. SES 12-58258 and thanks the Editor, a referee, and Ji Hyung Lee for helpful comments on the original version of the paper.

> *"There is no way to predict the price of stocks and bonds over the next few days or weeks. But it is quite possible to foresee the broad course of these prices over longer periods such as the next three to five years."* Press Release for the Prize in Economic Sciences, Royal Swedish Academy of Sciences, October 2013.

1 Introduction

Prediction is a central activity of econometics. From a practical standpoint, it is one of the more useful activities of the discipline, especially when it includes projections that enable us to analyze the effects of policy changes or to take advantage of investment opportunities. Prediction is also the most visible feature of econometrics to the public at large and perhaps the most relevant to the standing of the economics profession. The vulnerabilities of the profession in this dimension were thrown starkly into evidence when in November 2008 Queen Elizabeth asked a public gathering of economists at the London School of Economics why no one foresaw the financial crisis and global turmoil in financial markets. This question summarized ongoing reflection about the crisis in the public arena as well as debates that erupted within the economics and finance professions about the relevance of much macroeconomic theory and financial modeling.

In a subsequent show of faith in the discipline and concomitant with the subsequent recovery from the Great Recession that followed the financial crisis, in 2013 the Royal Swedish Academy awarded the Nobel Prize in economics to the 2013 Laureates for their work on the empirical analysis of asset pricing. The press release of the Academy cited in the headnote focused on the "surprising and contraditory" findings on predictability: financial asset prices are essentially unpredictable in the short run, but may be predictable in the long run. How can that be possible? The standard answer is simple: short-run volatility effectively masks long-term movements, which special econometric methods are needed to detect. But how successful are these methods. And would Queen Elizabeth be satisfied with such a response in the face of the substantial impact inflicted by the GFC on investors, the financial industry, the solvency of nation states, and ultimately the taxpayer?

Econometric forecasting now commands a vast literature, with dedicated journals and tentacles of empirical applications that stretch across the modern social and business sciences, as well as highly specific literatures in some areas like macroeconomics and finance. The extensive work on stock market forecasting formally began with the establishment by Alfred Cowles of the Cowles Commission in 1932 with the express purpose articulated in its *Articles of Incorporation* "to advance the scientific study and development .... of economic theory in its relation to mathematics and statistics" (Christ, 1952). Primary initial functions, besides collecting data and the formulation of indexes, were to analyze whether stock market forecasters could forecast successfully, and to provide a scientific foundation for that activity. In early research on the first question, as the header to this article attests, Cowles (1933, 1944) failed to find any evidence of ability to predict the stock market successfully, a conclusion that appears to be at variance with the later view of the Royal Swedish Academy. Financial markets, trading mechanisms, regulatory structures, global communications, corporations, and the world economy underlying stocks and bonds have undergone tectonic changes in size and complexity since Cowles (1933) initiated this line of analysis. With all these developments combined with enormous growth in the mutual fund industry, one may have expected markets to have become even more 'efficient' in their operations and correspondingly less 'predictable' in the intervening period. But econometric methodologies of detection have also changed substantially in the intervening decades. Moreover, electronic monitoring of market transactions on a tick by tick basis produces massive quantities of data that are now available for empirical work to analyze market behavior and trends.

Modern research on stock market predictability often involves trawling through these vast data sets hunting for predictive agents and extensive regression running with a mix of stationary and nonstationary time series and nonlinear functions of the raw data such as price/earnings and dividend/earnings ratios. The findings from such research has turned out to be somewhat ambiguous and baffling, as recently argued in the work of Welch and Goyal (2007). The quantities being estimated - commonly, the slope coefficients in the regressions - are frequently small (consonant with the Academy's statement concerning near martingale behavior in the short run) and the explanatory power of the regressions is generally low (consonant with Cowles's failure to find any evidence that forecasters are able to forecast financial markets successfully). Of course, big rare events - like the recent financial crisis – are often regarded as unpredicable in terms of precise timing and their specific form, but are now recognized as an inevitable feature of large complex systems. The theory of self organized criticality of complex physical and social systems, originating in the statistical physics work of Bak, Tang and Wiesenfeld (1987), provides strong arguments for the natural occurrence of such phenomena that involve periodic accumulation and collapse, just as in the metaphor of intermittent avalanches that occur naturally in a slowly accumulating sandpile.

Notwithstanding these qualifications, financial theory is remarkably silent on the issue of reconciling short run martingale behaviour with predictable long run behavior, begging questions of how long is necessary to achieve predictability, which are the key predictors or relevant predictive factors, and whether the predictive models are logically sound in terms of their time series properties. In consequence, much of the work in this field, as acknowledged by the Royal Swedish Academy, is empirically driven econometric research. That research has many pitfalls, as we will discuss. But it also offers many possibilities, including the use of new methodologies for inference.

The aim of the present paper is to overview certain aspects of this rapidly growing field of research. The paper studies some of the pitfalls in predictive regression while at the same time exploring some of the options made possible by recent econometric methodology. In part, therefore, the paper involves a review of existing methods combined with some recent research that is opening up new possibilities for empirical work.

The remainder of the paper is organized as follows. Section 2 examines some of the endemic pitfalls in predictive regression, covering new as well as commonly cited issues. We discuss some of the possibilities now available for addressing or bypassing these pitfalls, including some methods that are presented here for the first time. Section 3 explores some recent developments that have opened up new options in linear, nonparametric and quantile regression methods. Some concluding comments are given in Section 4.

2 Pitfalls and Possibilities

We examine a variety of problems encountered in the use of existing methods of predictive regression. Past literature in the field has acknowledged the most common problems and sought to find ways around these difficulties, as discussed in Phillips and Lee (2013). We start by briefly reviewing these issues, most of which stem from endogeneity problems and the fact that many commonly used explanatory regressors involve time series with varying, unknown degrees of persistence. Options for dealing with these difficulties or attenuating their impact on inference are discussed in each case.

For our discussion it is convenient to use the basic linear predictive model that is commonly formulated in triangular system format following Phillips (1991) as

$$y_t = \beta' x_{t-1} + u_{0t}, \tag{2.1}$$

$$x_t = \rho x_{t-1} + u_{xt}. \tag{2.2}$$

where the focus of attention is on the prediction of some scalar time series y_t given past information embodied in a set of regressors x_{t-1}. Since in much practical work like stock market or foreign exchange return forecasting the scalar dependent variable y_t has time series behavior that is close to a martingale difference sequence (mds) it is conventional to assume that the equation error u_{0t} is an mds with $\mathbb{E}\left(u_{0t}^2|\mathcal{F}_{t-1}\right) = \sigma_{00}$ *a.s.*, where $\mathcal{F}_t = \sigma\left(u_t, u_{t-1}, ...\right)$ is the natural filtration associated with the driver innovations $u_t = \left(u_{0t}, u_{xt}'\right)'$. This condition can readily be extended to allow for conditional heterogeneity, as will often be relevant when the observed time series are asset returns, but such extensions are not needed in explaining the primary pitfalls inherent in the predictive regression framework. For the purpose of the following discussion, let $(u_t, \mathcal{F}_t)$ be an mds with

$$\mathbb{E}\left(u_t u_t'|\mathcal{F}_{t-1}\right) = \begin{bmatrix} \sigma_{00} & \sigma_{0x} \\ \sigma_{x0} & \Sigma_{xx} \end{bmatrix} =: \Sigma,$$

which allows for x_t to be a vector of potential regressors useful in forecasting y_t and accommodates contemporaneous correlation between the components of the model. Again, it is easy to extend this structure to permit temporal dependence, intercepts and localized drifts (for the latter see Phillips, Shi and Yu, 2014) but it is helpful to keep to this simple framework to expound ideas.

2.1 Bias

Applying least squares to (2.1), assuming x_t is scalar, and setting $u_{0.xt} = u_{0t} - \sigma_{0x}\Sigma_{xx}^{-1}u_{xt}$ the estimation error decomposes as

$$\hat{\beta} - \beta = \frac{\sum_{t=1}^n x_{t-1}u_{0.xt}}{\sum_{t=1}^n x_{t-1}^2} + \sigma_{0x}\Sigma_{xx}^{-1}\left(\hat{\rho} - \rho\right), \tag{2.3}$$

where $\hat{\rho} = \left(\sum_{t=1}^n x_{t-1}^2\right)^{-1}\sum_{t=1}^n x_{t-1}x_t$. Taking expectations, we have

$$\mathbb{E}\left(\hat{\beta} - \beta\right) = \sigma_{0x}\Sigma_{xx}^{-1}\mathbb{E}\left(\hat{\rho} - \rho\right) = \sigma_{0x}\Sigma_{xx}^{-1}B_n\left(\rho\right) =: C_n\left(\Sigma, \rho\right)$$

where the autoregressive bias function $B_n(\rho) = \mathbb{E}(\hat{\rho} - \rho)$ depends only on ρ and n. An exact formula for $B_n(\rho)$ under Gaussianity is given in Phillips (2012) together with the following complete set of asymptotic expansions that hold for large n

$$B_n(\rho) = \begin{cases} -\frac{2\rho}{n} + O\left(n^{-2}\right) & |\rho| < 1 \\ -\frac{1.7814}{n} + O\left(n^{-2}\right) & \rho = 1 \\ -\frac{g(c)}{n} + O\left(n^{-2}\right) & \rho = 1 + \frac{c}{n} \\ O\left(\frac{1}{|\rho|^n}\right) & |\rho| > 1 \end{cases}, \tag{2.4}$$

where $g(c)$ is a continuous function of c whose analytic form is given in Phillips (2012). While in all cases the bias is negative for $\rho > 0$, the formulae show the discontinuities that occur in the bias expansions upon moving from a stationary regressor ($|\rho| < 1$) to a unit root (UR) regressor ($\rho = 1$), and through to an explosive regressor ($\rho > 1$). The local unit root (LUR) case with $\rho = 1 + \frac{c}{n}$ involves a continuous function $g(c)$ which has the property that $\lim_{c \to 0} g(c) = 1.7814$, $\lim_{c = o(n), c \to -\infty} g(c) = 2$, and $\lim_{c \to \infty} g(c) = 0$, which partially assists in bridging the stationary, unit root, and explosive cases.

Useful though the formulae given in (2.4) are, practical robust bias correction using them is not possible because, when ρ is unknown, so too is the precise formula to implement. Although the parameter ρ may be consistently estimated (again typically with bias), the localizing coefficient c is not consistently estimable except in very special circumstances (Moon and Phillips, 2000, 2004; Phillips, Moon and Xiao, 2001). Hence, the suggestion has been made in several papers (Stambaugh, 1999; Kothari and Shanken, 1997; Amihud and Hurvich, 2004) to correct bias using the particular version of (2.4) that holds for $|\rho| < 1$, in which case the correction applies only under the assumed condition of stationarity. Similar problems are encountered with higher order expansions. Such procedures inevitably err and lead to further bias when the condition fails and the regressors display persistence, as is commonly the case in practical work. Pre-test methods that use an estimate of ρ prior to selecting the appropriate bias formula produce further difficulties because of the presence of pre-test bias.

These problems of parameter dependence and discontinuity continue to apply and are typically more complex in the case of multiple regressors with more unknown parameters in the dynamics for x_t. Even in models where there are very large numbers of regressors and there is a common autoregressive coefficient, the bias problems remain – just as they are present in dynamic panel regressions under least squares estimation (Hahn and Kuersteiner, 2002).

It is now known that indirect inference methods can successfully deal with bias in autoregressive estimation and are robust to stationary and nonstationary values of ρ (Phillips, 2012). These methods may be used in the present context even though the autoregressive coefficient bias is only implicit in the estimation of β, as is apparent in the simple decomposition formula (2.3). For instance, ρ may be estimated robustly and with virtually no bias in finite samples by the indirect inference estimator $\breve{\rho}$, as described in Phillips (2012). Then the predictive regression coefficient $\hat{\beta}$ may be bias corrected using the formula

$$\breve{\beta}^{(1)} = \hat{\beta} - C_n\left(\hat{\Sigma}, \breve{\rho}\right) = \hat{\beta} - \hat{\sigma}_{0x}\hat{\Sigma}_{xx}^{-1} B_n\left(\breve{\rho}\right), \tag{2.5}$$

where $B_n\left(\breve{\rho}\right)$ plugs the indirect inference estimate $\breve{\rho}$ into the exact bias formula $B_n\left(\rho\right) = \mathbb{E}\left(\hat{\rho} - \rho\right)$ obtained analytically as in Phillips (2014) under a Gaussian assumption or by simulation. To complete the calculation in (2.5), $\hat{\Sigma}$ is a consistent estimate of Σ based on the residuals $\breve{u}_t = \left(\hat{u}_{0t}, \breve{u}_{xt}\right)'$ from the predictive regression $\hat{u}_{0t} = y_t - \hat{\beta}x_t$ and $\breve{u}_{xt} = x_t - \breve{\rho}x_{t-1}$ from the autoregressive equation fitted by using the indirect inference estimator $\breve{\rho}$. The process involved in (2.5) may be iterated to convergence using the scheme $\breve{\beta}^{j} = \hat{\beta} - C_n\left(\breve{\Sigma}^{j-1}, \breve{\rho}\right)$ with starting value $\breve{\Sigma}^{0} = \hat{\Sigma}$ and with $\breve{\Sigma}^{(j-1)}$ based on the $(j-1)$th iteration residuals $\breve{u}_{0t}^{(j-1)} = y_t - \breve{\beta}^{(j-1)}x_t$. Iteration then achieves compatibility between the resulting estimates, thereby delivering an indirect inference estimator $\breve{\beta}$ of β that satisfies the nonlinear equation

$$\breve{\beta} = \hat{\beta} - C_n\left(\breve{\Sigma}, \breve{\rho}\right) = \hat{\beta} - \breve{\sigma}_{0x}\breve{\Sigma}_{xx}^{-1} B_n\left(\breve{\rho}\right).$$

This procedure, which appears to be new, has the advantage that it directly accommodates the exact autoregressive bias in a robust way for all possible values of ρ and for the given sample size n. On the other hand, it applies rigorously only under Gaussianity and it does not generalize easily to more complex predictive regressions with multiple regressors in view of the additional difficulties involved in the implementation of indirect inference.

2.2 Nonstandard Inference

Much of the recent literature on predictive regression has focused on the use of explanatory variables in predictive regressions that have some degree of time series persistence such as dividend yields, book-to-price ratios, interest rates, or yield spreads. A natural first choice in analyzing such regressions was to develop methodology for near integrated or LUR time series that avoid insistence on the presence of a unit root in the data but allow for varying degrees of persistence that seem suited to many series intended to capture fundamentals. Accordingly, LUR asymptotics based on Chan and Wei (1987) and Phillips (1987) have come to play a large role in this literature, so that if $\rho = 1 + \frac{c}{n}$, we have by standard manipulations the following limit theory which uses the decomposition in Phillips (1989) - see also Cavanagh et al (1995)

$$\begin{aligned} n\left(\hat{\beta} - \beta\right) &= \frac{\frac{1}{n}\sum_{t=1}^{n} x_{t-1}u_{0t}}{\frac{1}{n^2}\sum_{t=1}^{n} x_{t-1}^2} \Longrightarrow \frac{\int_0^1 J_x^c(s)dB_0(s)}{\int_0^1 J_x^c(s)^2 dr} & (2.6) \\ &= \frac{\int_0^1 J_x^c(s)dB_{0.x}(s)}{\int_0^1 J_x^c(s)^2 dr} + \sigma_{0x}\Sigma_{xx}^{-1}\frac{\int_0^1 J_x^c(s)dB_x(s)}{\int_0^1 J_x^c(s)^2 dr} & (2.7) \\ &= \left[\int_0^1 J_x^c(s)^2 dr\right]^{-1/2}\xi_{0.x} + \sigma_{0x}\Sigma_{xx}^{-1}\frac{\int_0^1 J_x^c(s)dB_x(s)}{\int_0^1 J_x^c(s)^2 dr}. & (2.8) \end{aligned}$$

Here $J_x^c(s) = \int_0^s e^{c(s-p)}dB_x(p)$ is a linear diffusion, $B = (B_0, B_x')'$ is vector Brownian motion with variance matrix Σ, $B_{0.x} = B_0 - \sigma_{0x}\Sigma_{xx}^{-1}B_{0.x}$ and

$$\xi_{0.x} := \frac{\int_0^1 J_x^c(s)dB_{0.x}(s)}{\left(\int_0^1 J_x^c(s)^2 dr\right)^{1/2}} = N\left(0, \Sigma_{00.x}\right), \quad \Sigma_{00.x} = \sigma_{00} - \sigma_{0x}\Sigma_{xx}^{-1}\sigma_{x0}.$$

By construction, $B_{0.x}$ and $\xi_{0.x}$ are independent of B_x and therefore independent of the second term in (2.8). The source of the nonstandard limit theory in (2.6) and (2.8) therefore originates in endogeneity from non-zero correlation ($\sigma_{0x} \neq 0$) between the limit processes B_0 and B_x. Thus, although the regressor-error product element $x_{t-1}u_{0t}$ that appears in the numerator of (2.6) behaves nicely as a martingale difference sequence, the sample covariance $\frac{1}{n}\sum_{t=1}^{n} x_{t-1}u_{0t} \Rightarrow \int_0^1 J_x^c(s)dB_0(s)$ is non-zero and random, embodying limiting endogeneity effects from the correlation of the processes $J_x^c(s)$ and $B_0(s)$. More specifically, although the orthogonality condition $\mathbb{E}\left(x_{t-1}u_{0t}\right) = 0$ still holds, nonstationarity in the regressor x_{t-1} ensures that the limiting stochastic processes $J_x^c(s)$ and $B_0(s)$ are correlated when $\sigma_{0x} \neq 0$, which in turn leads to the endogeneity effect arising from the second term of (2.8). The end result is nonstandard limit theory behavior that is very different from the stationary ergodic case ($|\rho| < 1$) where the strong law $\frac{1}{n}\sum_{t=1}^{n} x_{t-1}u_{0t} \rightarrow_{a.s} \mathbb{E}\left(x_{t-1}u_{0t}\right) = 0$ holds and a standard CLT $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_{t-1}u_{0t} \Rightarrow N\left(0, \sigma_{00}\mathbb{E}\left(x_t^2\right)\right)$ applies with no endogeneity effect.

Indeed, when $|\rho| < 1$ (2.8) is replaced by $\sqrt{n}\left(\hat{\beta}-\beta\right) \Rightarrow N\left(0, \sigma_{00}\Sigma_{xx}^{-1}\right)$ and classical inferential tools of regression asymptotics apply. The discontinuity in the distributional asymptotics around $\rho = 1$ mirrors the discontinuity (2.4) in the asymptotic form of the bias function. And whilst the limit theory is classical when $|\rho| < 1$, there is inferential distortion from finite sample bias and skewness in such cases.

The primary difficulty empirical investigators face in conducting predictive inference is that there is uncertainty about the degree of persistence in the regressor. Standard methods fail in nonstationary cases and methods designed to treat LUR predictors often fail in stationary and certain mildly integrated (MI) cases, as does the popular Campbell and Yogo (2006) procedure (see Phillips, 2014). Moreover, multivariate regressors present further technical complexities and numerical complications for many methods, meaning that they are implementable in practice only in single predictor specifications. Differences in the persistence properties of the predictors cause additional difficulties and pre-test evaluation of the predictors for persistence induces pre-test bias. A particular difficulty associated with (2.8) is that the limit theory depends intimately on the localizing coefficient c, which is not consistently estimable, and therefore does not lead to a pivotal statistic for testing the predictability hypothesis $\mathbb{H}_0 : \beta = 0$.

Over the last two decades the econometric literature has struggled to find a satisfactory, practical way of dealing with these many complicating features of least squares predictive regression. We consider first the following two methods of dealing with uncertainty about the localizing coefficient c that grew out of the LUR limit theory.

(i) Bonferroni Methods

One mechanism for bypassing the dependence on c is to use Bonferroni bounds to find a confidence interval for β that incorporates confidence limits for c and therefore does not depend on a particular value of c. Given Σ (or a consistent estimator of Σ) such a confidence interval can be found by inversion of a suitable unit root test statistic under the LUR alternative $\rho = 1+\frac{c}{n}$ and taking upper and lower bounds over c, as suggested originally in Cavanagh, Elliott and Stock (1995) using ideas from Stock (1991) and confidence belt arguments from early statistical theory.

The approach was pursued systematically by Campbell and Yogo (2006; CY) in a form for predictive regression that quickly became influential and proved convenient for applied work. In brief, to construct a Bonferroni confidence interval (CI), the investigator constructs a $100\,(1-\alpha_1)\,\%$ CI for the localizing coefficient c, denoted as $CI_c(\alpha_1)$, by the test inversion method of Stock (1991). Then, for each value of c in this confidence interval, a $100\,(1-\alpha_2)\,\%$ CI for β is constructed based on that value of c, which is denoted by $CI_{\beta|c}\,(\alpha_2)$. A CI that does not depend on c is then obtained as

$$CI_\beta\,(\alpha) = \bigcup_{c \in CI_c(\alpha_1)} CI_{\beta|c}\,(\alpha_2)\,,$$

and by Bonferroni's inequality, this CI has coverage probability of at least $100\,(1-\alpha_1-\alpha_2)\,\%$. The approach appears by construction to be conservative and numerical size often turns out to be much less than nominal size when ρ is close to unity but this is certainly not the case for stationary values of ρ far from unity. The approach also results in a biased test because there are local alternatives for which power is less than nominal size. The approach is confined to the case of a scalar regressor and extensions to multiple regressors are impractical because of the need to cope with multiple localizing coefficients.

With a regressor whose autoregressive root is very close to unity, the CY approach does control size and has power for sizeable departures from the null. As a result, the method has been frequently employed in the applied literature. However, the method fails badly for values of ρ that approach the stationary region. The explanation for this failure is that the confidence intervals for c that are used in the procedure turn out to be invalid and are seriously biased asymptotically when the true value of ρ is stationary (Phillips, 2014). This failure of uniformity in the approach leads to poor performance in the CY predictive regression tests and CIs that are based on Bonferroni methods using LUR asymptotics when $|\rho| < 1$. Figure 1 (from Phillips, 2014) shows that CY confidence intervals have very poor coverage probabilities in the stationary case – in fact only a very small range of values of ρ deliver confidence intervals with close to nominal coverage and these values are clearly sensitive to the degree of endogeneity in the system as measured by the error correlation $r_{0x} = \frac{\sigma_{0x}}{(\sigma_{00}\sigma_{xx})^{1/2}}$. As $n \to \infty$ when $|\rho| < 1$, the CIs have zero coverage probability and false detection of predictability is therefore inevitable asymptotically under the null when the regressor is stationary. These results suggest substantial caution needs to be exercised in the use of these methods in practical work, where the degrees of persistence and endogeneity of the explanatory regressor are unknown. By contrast, the simple use of CIs based on stationary asymptotics leads – perhaps surprisingly – to a far greater degree of uniformity in ρ, where the 90% level and coverage probability of the stationary CIs are barely distinguishable on the scale of Figure 1.

As discussed in Phillips (2014), the CY Q test can be modified by changing the construction to employ a CI for ρ that is based on a centred test statistic for ρ such as the t test

$$t_{\hat{\rho},\rho} = \frac{\hat{\rho} - \rho}{\sigma_{\hat{\rho}}} \Longrightarrow \frac{\int_0^1 J_x^c dB_x}{\left(\int_0^1 J_x^c(s)^2 ds\right)^{1/2}} = \lambda_c, \text{ with } \sigma_{\hat{\rho}}^2 = \frac{n^{-1} \sum_{t=1}^n \left(x_t - \hat{\rho} x_{t-1}\right)^2}{\sum_{t=1}^n x_{t-1}^2},$$

rather than a unit root test (with $\rho = 1$). As Mikusheva (2007) shows, this construction leads to a uniformly valid CI for ρ under some general conditions. Moreover, from Phillips (1987, 2014), the limit variate of the centred statistic $\lambda_c \sim N(0,1) + O_p\left(|c|^{-1/2}\right)$ as $c \to -\infty$ and under these conditions the induced CI for ρ is approximately $[\rho_L, \rho_U] = \left\{\hat{\rho} - z_{\alpha_1/2}\sigma_{\hat{\rho}}, \hat{\rho} + z_{\alpha_1/2}\sigma_{\hat{\rho}}\right\}$ for a nominal level α_1 test, which is asymptotically valid for $|\rho| < 1$ matching the stationary asymptotics. The corresponding CI for β has coverage probability that is at least $100\,(1 - \alpha_2 - \alpha_1)\,\%$ by Bonferroni. Hence, use of the centred test statistic $t_{\hat{\rho},\rho}$ for ρ leads to a robust interval for which the Bonferroni bound holds and this construction of the CI avoids the zero coverage probability in the stationary case of the CY interval based on the Q test. Computation of this modified interval requires the use of confidence belts for ρ based on the centred statistic $t_{\hat{\rho},\rho}$. While valid, this modification of the method still encounters an impassable numerical obstacle for multivariate x_t with multiple nuisance parameters arising from the localizing coefficients associated with each individual regressor.

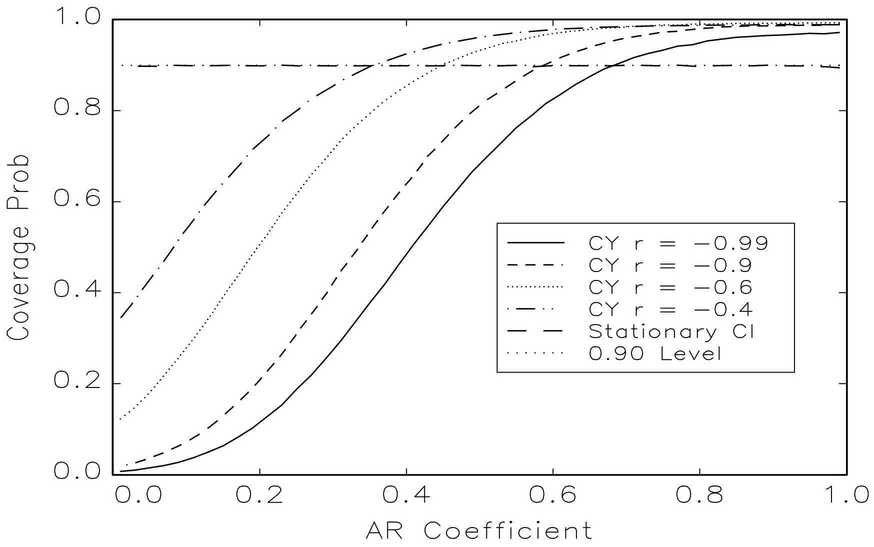

Fig. 1 (from Phillips, 2014): Coverage probabilities of Campbell-Yogo and stationary confidence intervals for the predictive regression coefficient β plotted against the autoregressive coefficient ρ of x_t, shown for various values of the endogeneity coefficient $r_{0x} = \frac{\sigma_{0x}}{(\sigma_{00}\sigma_{xx})^{1/2}}$. The nominal asymptotic level is 90%, sample size is $n = 200$, and the number of replications is $50,000$.

(ii) Test Statistic Conditioning

A second approach that is designed to avoid dependence on c involves test conditioning and was suggested by Jansson and Moreira (2006; JM). The idea uses the Gaussian likelihood and its asymptotic form to produce a conditional likelihood test for β in terms of sufficient statistics. Jointly sufficient statistics for (β, ρ) in (2.1) and (2.2) are used to constructed a conditional likelihood whose distribution does not depend on the localizing coefficient c and this likelihood is used to obtain test critical values. Tests based on this conditional likelihood ratio approach attain optimality within a certain class of conditional, similar tests.

In principle this method has attractive features. It is likelihood based and has associated optimality properties. Nonetheless, practical experience with the JM test has been disappointing, as simulation evidence from many studies attest (e.g., Chen and Deo, 2009; Kasparis et al, 2015; Kostakis et al, 2015). In part, this is due to algorithmic complications arising from quadrature that is required for implementation, for which numerical difficulties have been encountered (e.g. Kasparis, et al, 2015). In part, also, simulation experience reveals that finite sample performance of the JM test is erratic. Size and power are often bettered by other tests which have no claimed optimality properties – such as Campbell and Yogo (2006), Chen and Deo (2009), Kostakis et al (2015), Kasparis et al (2015), and Phillips and Chen (2014).

To illustrate the arguments that underlie the effects of conditioning, it is convenient to let the covariance matrix of the innovations $u_t = (u_{0t}, u_{xt})'$ in the system be written in standardized form as $\Sigma = \mathbb{E}\left(u_t u_t' | \mathcal{F}_{t-1}\right)$ with $\sigma_{00} = \Sigma_{xx} = 1$, and $\sigma_{0x} = r_{0x}$. Then, the functional law $n^{-1/2}\sum_{t=1}^{\lfloor n\cdot \rfloor} u_t \Rightarrow W(\cdot) = \left(W_0(\cdot), W_x(\cdot)\right)'$ holds, where W is vector Brownian motion (BM) with variance matrix Σ and $W_{0.x}(r) = W_0(r) - r_{0x}W_x(r) \equiv BM\left(1 - r_{0x}^2\right)$, which is independent of W_x. In what follows we will assume the correlation r_{0x} is known since consistent estimation of r_{0x} (or the covariance matrix Σ) does not present any difficulties or disturb the arguments concerning the dependence of test statistics on c. We assume for the moment that $|r_{0x}| < 1$ and later examine the strong endogeneity case where $|r_{0x}| = 1$.

Set $u_{0.xt} = u_{0t} - r_{0x}u_{xt}$ and rewrite the model as

$$\begin{aligned} y_t &= \left(\beta - r_{0x}\frac{c}{n}\right)x_{t-1} + r_{0x}\Delta x_t + u_{0.xt}, \\ x_t &= \rho_n x_{t-1} + u_{xt}, \quad \text{with } \rho_n = 1 + \frac{c}{n}. \end{aligned}$$

Let $\beta = b/n$, so that β is local to zero. Since $\mathbb{E}u_{0.xt}^2 = 1 - r_{0x}^2$, the Gaussian log likelihood function up to a constant is

$$\ell_n(b,c) = -\frac{1}{2\left(1-r_{0x}^2\right)}\sum_{t=1}^{n}\left\{y_t - r_{0x}\Delta x_t - \frac{b - r_{0x}c}{n}x_{t-1}\right\}^2 - \frac{1}{2}\sum_{t=1}^{n}\left(\Delta x_t - \frac{c}{n}x_{t-1}\right)^2, \tag{2.9}$$

from which the log likelihood ratio $\Lambda_n(b,c) := \ell_n(b,c) - \ell_n(0,0)$ is

$$\Lambda_n(b,c) = \frac{b}{\left(1-r_{0x}^2\right)}S_\beta + cS_\gamma - \frac{1}{2}\left\{\frac{b^2 - 2r_{0x}bc + c^2}{\left(1-r_{0x}^2\right)}\right\}S_{\gamma\gamma}, \tag{2.10}$$

where $S_\beta = \frac{1}{n}\sum_{t=1}^{n} x_{t-1}\left(y_t - r_{0x}\Delta x_t\right)$, $S_{\gamma\gamma} = \frac{1}{n^2}\sum_{t=1}^{n} x_{t-1}^2$, $S_\gamma = \frac{1}{n}\sum_{t=1}^{n}\Delta x_t x_{t-1} - r_{0x}S_\beta$.

Since $y_t - r_{0x}\Delta x_t = \frac{(b-r_{0x}c)}{n}x_{t-1} + u_{0.xt}$, standard limit theory (Phillips, 1987a & b) gives

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n} x_{t-1}\left(y_t - r_{0x}\Delta x_t\right) &\Rightarrow \int_0^1 J_x^c dW_{0.x} + (b - r_{0x}c)\int_0^1 J_x^{c2}, \\ \frac{c}{n}\sum_{t=1}^{n}\Delta x_t x_{t-1} &\Rightarrow \int_0^1 J_x^c dJ_x^c,\ \frac{1}{n^2}\sum_{t=1}^{n} x_{t-1}^2 \Rightarrow \int_0^1 J_x^{c2}.\end{aligned}$$

From these limits and (2.10) we deduce that

$$\Lambda_n(b,c) \Rightarrow \frac{b}{1-r_{0x}^2}\mathcal{R}_\beta + c\mathcal{R}_\gamma - \frac{1}{2}\left\{\frac{(b-r_{0x}c)^2}{1-r_{0x}^2} + c^2\right\}\mathcal{R}_{\gamma\gamma} =: \mathcal{L}(b,c) \tag{2.11}$$

where

$$\begin{aligned}\mathcal{R}_\beta(b,c) \ &: = \int_0^1 J_x^c dW_{0.x} + (b - r_{0x}c)\int_0^1 J_x^{c2}, \\ \mathcal{R}_\gamma(b,c) \ &: = \int_0^1 J_x^c dJ_x^c - \frac{r_{0x}}{1-r_{0x}^2}\mathcal{R}_\beta, \quad \text{and } \mathcal{R}_{\gamma\gamma} := \mathcal{R}_{\gamma\gamma}(c) := \int_0^1 J_x^{c2},\end{aligned}$$

analogous to equation (17) in JM.[1]

Under the null hypothesis of no predictability ($\beta = b = 0$) the limiting log likelihood ratio is $\mathcal{L}(0,c) = c\mathcal{R}_\gamma(0,c) - \frac{c^2}{2\left(1-r_{0x}^2\right)}\mathcal{R}_{\gamma\gamma}$, so that all information about c in the limit is contained in the quantities $\left(\mathcal{R}_\gamma(0,c), \mathcal{R}_{\gamma\gamma}\right)$. The JM test is based on the conditional distribution of the statistic $\mathcal{R}_\beta$ given $\left(\mathcal{R}_\gamma, \mathcal{R}_{\gamma\gamma}\right)$. Following in the same way as JM (2006, Lemma 4), the joint density of $\mathcal{R} = \left(\mathcal{R}_\beta, \mathcal{R}_\gamma, \mathcal{R}_{\gamma\gamma}\right)$ at $r = \left(r_\beta, r_\gamma, r_{\gamma\gamma}\right)$ has the form

$$f_{\mathcal{R}}(r; b, c) = K(b,c)\, f_{\mathcal{R}}^0(r)\exp\left[\frac{b}{1-r_{0x}^2}r_\beta + cr_\gamma - \frac{1}{2}\left\{\frac{(b-r_{0x}c)^2}{1-r_{0x}^2} + c^2\right\}r_{\gamma\gamma}\right],$$

where $f_{\mathcal{R}}^0(r)$ is the joint density of $\mathcal{R}$ for $(b,c) = (0,0)$ and is therefore independent of c.

---

[1] The limiting log likelihood ratio $\mathcal{L}(b,c)$ in (2.11) is in the same form as JM's equation (17) and relates to it by: (i) rescaling b with $\left(1-r_{0x}^2\right)^{1/2}$, since JM parameterize β as $\beta = b\left(1-r_{0x}^2\right)^{1/2}/n$; and (ii) using the alternative definition $\mathcal{R}_\beta = \int J_x^c d\bar{W}_{0.x}$ where $\bar{W}_{0.x}(r) = \left(1-r_{0x}^2\right)^{-1/2} W_{0.x}(r) \equiv BM(1)$. Moreover, with the absence of an intercept in the predictive regression, JM's component $\mathcal{R}_{\beta\beta}$ is identical to $\mathcal{R}_{\gamma\gamma}$, so the minimal asymptotic sufficient statistic for $a = (b,c)$ is simply $\mathcal{R} = \left(\mathcal{R}_\beta, \mathcal{R}_\gamma, \mathcal{R}_{\gamma\gamma}\right)$.

The conditional distribution of $\mathcal{R}_\beta$ given $(\mathcal{R}_\gamma, \mathcal{R}_{\gamma\gamma})$ is

$$f_{\mathcal{R}}\left(r_\beta | r_\gamma, r_{\gamma\gamma}; b\right) = \frac{f^0_{\mathcal{R}}\left(r_\beta, r_\gamma, r_{\gamma\gamma}\right) e^{b r_\beta}}{\int f^0_{\mathcal{R}}\left(r\right) \exp\left(b r_\beta\right) dr_\beta}, \tag{2.12}$$

which is also independent of c. Hence the conditional distribution of $\mathcal{R}_\beta$ given $(\mathcal{R}_\gamma, \mathcal{R}_{\gamma\gamma})$, being independent of c, can be used to produce a similar test of $\mathbb{H}_0 : \beta = b = 0$. Note that under the null we have

$$f_{\mathcal{R}}\left(r_\beta | r_\gamma, r_{\gamma\gamma}; b = 0\right) = \frac{f^0_{\mathcal{R}}\left(r_\beta, r_\gamma, r_{\gamma\gamma}\right)}{\int f^0_{\mathcal{R}}\left(r\right) dr_\beta},$$

and so p-values for the observed $\hat{\mathcal{R}}_\beta$ are computed under $\mathbb{H}_0$ using

$$\frac{\int_{\hat{\mathcal{R}}_\beta}^\infty f^0_{\mathcal{R}}\left(r_\beta, r_\gamma, r_{\gamma\gamma}\right) dr_\beta}{\int_{-\infty}^\infty f^0_{\mathcal{R}}\left(r\right) dr_\beta}.$$

It is of interest to determine the effects of conditioning directly on the variate $\mathcal{R}_\beta$. Observe that

$$\begin{aligned} \mathcal{R}_\beta &= \mathcal{R}_\beta\left(b, c\right) = \int_0^1 J^c_x(r) dW_{0.x}(r) + \left(b - r_{0x} c\right) \int_0^1 J^c_x(r)^2 dr \\ &= \left(\int_0^1 J^c_x(r)^2 dr\right)^{1/2} \xi_{0.x} + \left(b - r_{0x} c\right) \int_0^1 J^c_x(r)^2 dr, \end{aligned}$$

where $\xi_{0.x} \sim_d N\left(0, 1 - r^2_{0x}\right)$ is independent of $\int_0^1 J^c_x(r)^2 dr$, $\int_0^1 J^c_x(r) dW_x(r)$, and c. It follows that $\mathcal{R}_\beta\left(b, c\right)$ has the following conditional normal distribution, conditional on $\mathcal{R}_{\gamma\gamma}\left(c\right)$,

$$\mathcal{R}_\beta\left(b, c\right)|_{\mathcal{R}_{\gamma\gamma}(c)} \sim_d N\left(\left(b - r_{0x} c\right) \mathcal{R}_{\gamma\gamma}, \left(1 - r^2_{0x}\right) \mathcal{R}_{\gamma\gamma}\right),$$

which is not independent of c because the mean relies on $\left(b - r_{0x} c\right) \mathcal{R}_{\gamma\gamma}$ whose factor $\left(b - r_{0x} c\right)$ depends on c, at least when $r_{0x} \neq 0$. Next, observe that

$$\mathcal{R}_\gamma = \int_0^1 J^c_x(r) dJ^c_x\left(r\right) - \frac{r_{0x}}{1 - r^2_{0x}} \mathcal{R}_\beta = c \int_0^1 J^c_x(r)^2 dr + \int_0^1 J^c_x(r) dW_x\left(r\right) - \frac{r_{0x}}{1 - r^2_{0x}} \mathcal{R}_\beta,$$

so that for $r_{0x} \neq 0$

$$\mathcal{R}_\beta = \frac{1-r_{0x}^2}{r_{0x}} \left\{ c\mathcal{R}_{\gamma\gamma} + \int_0^1 J_x^c(r) dW_x(r) - \mathcal{R}_\gamma \right\} = \frac{1-r_{0x}^2}{r_{0x}} \left\{ \frac{1}{2} \left\{ J_x^c(1)^2 - 1 \right\} - \mathcal{R}_\gamma \right\},$$

since from Philllips (1987b)

$$\int_0^1 J_x^c(r) dW_x(r) = \frac{1}{2} \left\{ J_x^c(1)^2 - 1 \right\} - c \int_0^1 J_x^c(r)^2 dr = \frac{1}{2} \left\{ J_x^c(1)^2 - 1 \right\} - c\mathcal{R}_{\gamma\gamma}.$$

It follows that conditioning $\mathcal{R}_\beta$ on $\mathcal{R}_{\gamma\gamma}$ and $\mathcal{R}_\gamma$ we have when $r_{0x} \neq 0$

$$\mathcal{R}_\beta |_{\mathcal{R}_{\gamma\gamma}, \mathcal{R}_\gamma} = \frac{1-r_{0x}^2}{2r_{0x}} \left\{ J_x^c(1)^2 - 1 \right\} |_{\mathcal{R}_{\gamma\gamma}(c), \mathcal{R}_\gamma} - \mathcal{R}_\gamma. \tag{2.13}$$

Now $J_x^c(1) = \int_0^1 e^{c(1-s)} dW(s) \sim_d N\left(0, \int_0^1 e^{2c(1-s)} ds\right) = N\left(0, \frac{1-e^{2c}}{-2c}\right)$ and $J_x^c(1)^2 = \frac{1-e^{2c}}{-2c}\xi^2$, with $\xi \sim_d N(0,1)$, has a scale factor $\frac{1-e^{2c}}{-2c}$ that is dependent on c. So if the conditional distribution of $\mathcal{R}_\beta$ given $(\mathcal{R}_{\gamma\gamma}(c), \mathcal{R}_\gamma(c))$ is independent of c according to (2.12), then the conditional distribution of $J_x^c(1)^2$ given $(\mathcal{R}_{\gamma\gamma}(c), \mathcal{R}_\gamma(c))$ must also be independent on c, a result that the author has not been able to verify.

An interesting feature of the null case with $\beta = b = 0$ that has not been noticed in the literature is that data on y_t affect the limit distribution of the maximum likelihood estimator, whereas this is not the case under the alternative. In particular, since the limiting log likelihood ratio for $b = 0$ is $\mathcal{L}(0,c) = c\mathcal{R}_\gamma(0,c) - \frac{c^2}{2\left(1-r_{0x}^2\right)}\mathcal{R}_{\gamma\gamma}(c)$, it follows that the limiting distribution of the restricted maximum likelihood estimator is simply

$$\tilde{c} = \frac{\mathcal{R}_\gamma(0,c)}{\mathcal{R}_{\gamma\gamma}(c)} = c + \frac{\int_0^1 J_c(r) dW_{x.0}(r)}{\int_0^1 J_c(r)^2}. \tag{2.14}$$

Under the alternative where $b \neq 0$, maximization of the limiting log likelihood jointly with respect to (b,c) yields the usual decomposition for $\hat{b}$ (c.f., (2.7) above)

$$\hat{b} - b = r_{0x}(\hat{c} - c) + \frac{\int_0^1 J_c(r) dW_{0.x}(r)}{\mathcal{R}_{\gamma\gamma}},$$

with the following limit for the unrestricted estimate of c

$$\hat{c} = \frac{\int J_c(r)dJ_c\left(r\right)}{\mathcal{R}_{\gamma\gamma}} = c + \frac{\int_0^1 J_c(r)dW_x\left(r\right)}{\int_0^1 J_c(r)^2}, \tag{2.15}$$

which differs from the restricted case (2.14). Importantly, the limit (2.15) corresponds exactly to single equation autoregressive LUR limit theory. So, in this case, the information set relevant to the estimation of c is the pair $\left(\int J_c(r)dJ_c\left(r\right), \mathcal{R}_{\gamma\gamma}\right)$ as in single equation autoregressive estimation, which does not involve any information from the prediction equation. By contrast, in the restricted case the limit theory in the numerator of $\tilde{c} - c$ is the stochastic integral $\int_0^1 J_c(r)dW_{x.0}\left(r\right)$ taken with respect to the conditional Brownian motion $W_{x.0}\left(r\right) = W_x\left(r\right) - r_{0x}W_0\left(r\right) \equiv BM\left(1 - r_{0x}^2\right)$ whose variance is smaller than $W_x\left(r\right)$ for all $r_{0x} \neq 0$, i.e., for all cases where the prediction equation error u_{0t} is correlated with the autoregressive equation error u_{xt}.

The intuitive explanation for the reduction in variance under the null is simply that in the restricted case where $b = 0$ the prediction equation is $y_t = u_{0t}$, so that information on y_t may be used to reduce variance in the estimation of c. To see this, note that the autoregressive equation may in this case be written as

$$x_t = \rho_n x_{t-1} + u_{xt} = \rho_n x_{t-1} + r_{0x}y_t + u_{x.0t},$$

or equivalently $x_t - r_{0x}y_t = \rho_n x_{t-1} + u_{x.0t}$, showing that knowledge of y_t can be used to reduce the error variance in the autoregressive equation, thereby raising the signal to noise ratio, much like the case of autoregressive equations that include covariate regressors in the UR or LUR cases (c.f., Hansen, 1995). An extreme case occurs under strong endogeneity where $r_{0x} = 1$. Then $u_{0t} = u_{xt}$ *a.s.* and $u_{x.0t} = 0$ *a.s.*, so that now $x_t = \rho_n x_{t-1} + y_t$ *a.s.* and c is known directly from the data.

2.3 Quantile Predictive Regressions and Crossing Problems

In place of linear mean predictive regressions of the form (2.1) and (2.2), attention has recently been given to quantile regression formulations. These are useful, as in other applications of quantile methods, when interest focuses on specific parts of the distribution of the dependent variable y_t and there is reason to expect that the response function to driver variables may differ in different parts of the distribution. The approach seems particularly well suited to financial applications where the effects on asset returns may well differ depending on whether the impact of a certain driver is positive or negative. Stylized features such as heavy tails and asymmetric distributions suggest that predictability from some driver variables may be greater at certain quantiles than at others, indicating the possible advantages of a quantile structure in model formulation. Examples of work of this type include Xiao (2009), Cenesizoglu and Timmerman (2008), Maynard et al (2011), and Lee (2015).

The quantile regression predictive model follows the standard formulation

$$\begin{aligned} Q_{y_t}(\tau|\mathcal{F}_{t-1}) &= \beta(\tau)x_{t-1} + F_{u_0}^{-1}(\tau), \\ x_t &= \rho x_{t-1} + u_{xt} \end{aligned} \tag{2.16}$$

where u_{0t} is assumed to be *iid* with *cdf* given by F_{u_0} so that $F_{u_0}^{-1}(\tau)$ is the unconditional τ-quantile of u_{0t}, and $Q_{y_t}(\tau|\mathcal{F}_{t-1})$ is the conditional τth quantile function of the distribution of y_t given the past information in $\mathcal{F}_{t-1}$. The regressor in (2.16) follows (2.2) and the model has error input $u_t = (u_{0t}, u_{xt})'$ satisfying the same conditions as in (2.1) and (2.2). In (2.16), the slope coefficient is allowed to vary according to the quantile τ and the conditional quantile formulation (2.16) is assumed to hold almost surely.

More generally, as in Maynard et al (2012) and Lee (2015), we can model the conditional quantile of the error term u_{0t} in a general way such that the predictive quantile regression model has the form

$$Q_{y_t}(\tau|\mathcal{F}_{t-1}) = \alpha(\tau) + \beta(\tau)x_{t-1}, \tag{2.17}$$

which allows the intercept and influence of the regressor x_{t-1} to be heterogenous across quantiles of y_t. This model accommodates conditional heterogeneity. For instance, as discussed in Maynard et al (2012), suppose the generating model for y_t has the form

$$y_t = \alpha_0 + \beta_0 x_{t-1} + (\alpha_1 + \beta_1 x_{t-1})u_{0t},$$

and suppose the conditional distribution of u_{0t} is $F_{u_0,t-1}(\cdot) = P(u_{0t} < \cdot|\mathcal{F}_{t-1}) = F_{u_0}(\cdot)$. Then (2.17) holds with

$$\beta(\tau) = \beta_0 + \beta_1 F_{u_0}^{-1}(\tau), \quad \text{and } \alpha(\tau) = \alpha_0 + \alpha_1 F_{u_0}^{-1}(\tau).$$

In the general case, we may define the predictive quantile function $Q_{y_t}(\tau|\mathcal{F}_{t-1}) =$

$\alpha(\tau) + \beta(\tau) x_{t-1}$ and the implied innovation is

$$u_{0t\tau} = u_{0t} - F_{u_0,t-1}^{-1}(\tau) = y_t - \alpha(\tau) - \beta(\tau) x_{t-1},$$

so that $Q_{u_{0t\tau}}(\tau|\mathcal{F}_{t-1}) = 0$ and $\psi_\tau(u_{0t\tau}) = \tau - \mathbf{1}\left\{u_{0t\tau} < F_{u_0}^{-1}(\tau)\right\}$. Then $E[\psi_\tau(u_{0t\tau})|\mathcal{F}_{t-1}] = 0$ and the variance of the indicator random variable $\mathbf{1}\left\{u_{0t\tau} < F_{u_0}^{-1}(\tau)\right\}$ is $\tau(1-\tau)$.

Crossing Probabilities in Quantile Formulations

As is well known from conventional quantile theory, the regression formulation (2.16) fails to be consistent across quantiles τ_1 and τ_2 when the natural quantile ordering ($Q_{y_t}(\tau_2|\mathcal{F}_{t-1}) > Q_{y_t}(\tau_1|\mathcal{F}_{t-1})$ for $\tau_2 > \tau_1$) is reversed by virtue of the posited regression formulation. Such reversals, or quantile crossings, occur whenever

$$\{\beta(\tau_2) - \beta(\tau_1)\} x_{t-1} + \left\{F_{u_0}^{-1}(\tau_2) - F_{u_0}^{-1}(\tau_1)\right\} < 0,$$

that is for

$$x_{t-1} < \frac{F_{u_0}^{-1}(\tau_2) - F_{u_0}^{-1}(\tau_1)}{\beta(\tau_2) - \beta(\tau_1)}, \quad \text{if } \beta(\tau_2) - \beta(\tau_1) > 0. \tag{2.18}$$

Reversals of this type signal misspecification in the quantile regression formulation, since for the given parameterization the model cannot be valid almost surely when there is a positive probability of a reversal such as (2.18). For a stationary time series predictor x_t with invariant measure P_x (e.g., when $x_t \sim_d N\left(0, \frac{\sigma_{xx}}{1-\rho^2}\right)$), the probability of such reversals is

$$P_x\left(\frac{F_{u_0}^{-1}(\tau_2) - F_{u_0}^{-1}(\tau_1)}{\beta(\tau_2) - \beta(\tau_1)}\right) \text{ for } \beta(\tau_2) > \beta(\tau_1). \tag{2.19}$$

If $\rho = 1$ and the generating mechanism for x_t is a unit root model, then there is no invariant measure in view of the nonstationarity of x_t. Instead, we can write the quantile crossing frequency as

$$n^{-1}\sum_{t=1}^{n} 1\left\{\frac{x_{t-1}}{\sqrt{n}} < \frac{1}{\sqrt{n}}\frac{F_{u_0}^{-1}(\tau_2) - F_{u_0}^{-1}(\tau_1)}{\beta(\tau_2) - \beta(\tau_1)}\right\}.$$

By standard tools of limit theory for nonlinear functions of integrated processes (Park and

Phillips, 1999, 2000, 2001) we find that the limiting form of this crossing frequency is

$$n^{-1}\sum_{t=1}^{n}\mathbf{1}\left\{\frac{x_{t-1}}{\sqrt{n}}<\frac{1}{\sqrt{n}}\frac{F_{u0}^{-1}(\tau_2)-F_{u0}^{-1}(\tau_1)}{\beta(\tau_2)-\beta(\tau_1)}\right\}\Rightarrow\int_0^1\mathbf{1}\{B_x(r)<0\}\,dr, \tag{2.20}$$

which is the amount of time the Brownian motion B_x spends below the horizontal axis over the interval $[0,1]$. Importantly, since scale does not matter within the indicator function, we have $\int_0^1\mathbf{1}\{B_x(r)<0\}\,dr=\int_0^1\mathbf{1}\{W(r)<0\}\,dr$, which is the soujourn time for $r\in[0,1]$ of a standard Brownian motion W on the half line $(-\infty,0)$. The distribution of the limit variate (2.20) is well-known to be the arc-sine law with probability density $\frac{1}{\pi\sqrt{x(1-x)}}$ over the support $x\in(0,1)$, which is a Beta distribution with parameters $\left(\frac{1}{2},\frac{1}{2}\right)$. This distribution is $\cup$ shaped over its support with asymptotes at the boundary points $\{0,1\}$ of the domain of definition. Thus, depending on the realization of the time series $\{x_t\}_1^n$, there is a far greater probability of either many crossings or few crossings in the quantile function.

Correspondingly, the probability of failure in the quantile regression formulation (2.16) differs significantly between stationary and nonstationary cases. The failure probability is fixed in the stationary case, is given explicitly by (2.19), and depends on the precise parameter values $(\tau_1,\tau_2,\beta(\tau_2)-\beta(\tau_1))$. In the unit root case, the failure frequency is not fixed but instead depends on the actual trajectory of $\{x_t\}$. In the limit, the failure frequency depends on the trajectory of the limiting Brownian motion B_x associated with the limit of the standardized process $X_n(\cdot)=\frac{x_{t=\lfloor n\cdot\rfloor}}{\sqrt{n}}$. Importantly, in this nonstationary case, the failure probability does not depend on the specific parameters $(\tau_1,\tau_2,\beta(\tau_2)-\beta(\tau_1))$, at least in the limit as $n\to\infty$. The form of the arc sine law limit theory for (2.20) shows that there will always be a high probability of quantile crossings, whatever the parameter values and functional dependence of the quantile slope coefficients $\beta(\tau)$, provided $\beta(\tau_2)\neq\beta(\tau_1)$. That is, provided the slope coefficient function $\beta(\tau)$ is non-constant, the quantile regression model is inevitably misspecified with high probability for unit root nonstationary regressors.

Similar findings apply in the case of an LUR predictor x_t with AR coefficient $\rho=1+\frac{c}{n}$. In place of (2.20) we then have

$$n^{-1}\sum_{t=1}^{n}\mathbf{1}\left\{\frac{x_{t-1}}{\sqrt{n}}<\frac{1}{\sqrt{n}}\frac{F_{u0}^{-1}(\tau_2)-F_{u0}^{-1}(\tau_1)}{\beta(\tau_2)-\beta(\tau_1)}\right\}\Rightarrow\int_0^1\mathbf{1}\{J_x^c(r)<0\}\,dr,$$

which is the soujourn time over $r\in[0,1]$ of a one dimensional standard diffusion $J_c(r)$

on the half line $(-\infty, 0)$. The distribution of the occupation time in this case and various other limiting stochastic process extensions have been extensively studied in the literature following Lamperti (1958).

Localized Validity of Quantile Formulations

These results suggest that (i) quantile predictive regressions require constant slope coefficients to assure full (i.e., almost sure) validity in specification, and (ii) the failure of consistency across quantiles is likely to be aggravated when persistent regressors are present leading to a high probability of reversals for some data trajectories. The requirement of constancy in the slope coefficient $\beta(\tau)$ across quantiles for the validity of quantile regression is highly restrictive and obviously defeats the primary purpose of quantile regression. Fortunately, the requirement may be relaxed by allowing for certain local departures of these coefficients from a constant value, as we now discuss.

Define the local to constant slope parameter $\beta(\tau) = \beta + \frac{b(\tau)}{d_n}$ where d_n is a sequence of positive numbers satisfying $d_n \to \infty$ as $n \to \infty$ and where $b(\tau)$ is a localizing coefficient function that may vary across quantiles over a domain such that $b(\tau) \in [b_L, b_U]$ for some finite b_L and $b_{U.}$ The local quantile predictive regression then has the (triangular array) form

$$Q_{y_t}(\tau|\mathcal{F}_{t-1}) = \left(\beta + \frac{b(\tau)}{d_n}\right) x_{t-1} + F_{u_0}^{-1}(\tau).$$

For this formulation, the condition for no reversals (no quantile crossing) is, for $\tau_2 > \tau_1$,

$$\left\{\frac{b(\tau_2) - b(\tau_1)}{d_n}\right\} x_{t-1} + \left\{F_{u_0}^{-1}(\tau_2) - F_{u_0}^{-1}(\tau_1)\right\} > 0.$$

Then, in the stationary regressor case where $|\rho| < 1$, the condition holds with probability approaching unity because $\frac{x_t}{d_n} \to_p 0$ for all $k_n \to \infty$. In the unit root case $\rho = 1$ we have

$$\sqrt{n}\left\{\frac{b(\tau_2) - b(\tau_1)}{d_n}\right\} \frac{x_{t-1}}{\sqrt{n}} + \left\{F_{u_0}^{-1}(\tau_2) - F_{u_0}^{-1}(\tau_1)\right\} > 0, \tag{2.21}$$

and this condition then holds with probability approaching unity provided $\frac{\sqrt{n}}{d_n} \to 0$ as $n \to \infty$. The same condition holds in the LUR case with $\rho = 1 + \frac{c}{n}$. Further, if x_t is mildly integrated in the sense that $\rho = 1 + \frac{c}{k_n}$ with $k_n \to \infty$ and $c < 0$ (see (3.1) and the discussion in section 3.1 below), then the 'no crossings' result continues to hold with probability approaching unity provided $\frac{\sqrt{k_n}}{d_n} \to 0$ as $n \to \infty$. Thus, in certain local neighborhoods of the parameter space that shrink as $n \to \infty$ to a constant slope β fast enough relative to the extent of the departure of ρ from unity, the probability of quantile reversals can be eliminated asymptotically, under suitable conditions on the rate, for stationary, nonstationary, and various local to unity and mildly integrated time series regressors.

Such cases allow validly for small local departures from constancy in the quantile regression coefficients $\beta(\tau)$ with some prospect of estimating the coefficients that distinguish the quantile slope coordinates. In particular, when $\rho = 1$, we may use fully modified quantile regression (Xiao, 2009) to estimate the slope coefficients $\beta(\tau)$. This approach is explored in Xiao's paper and applies fully modified methods from linear cointegrating regression theory for unit root regressors (Phillips and Hansen, 1990) within the quantile regression model to deliver an estimate $\hat{\beta}^{+}(\tau)$ which has the following mixed normal (MN) asymptotic distribution

$$n\left(\hat{\beta}^{+}(\tau) - \beta(\tau)\right) \Rightarrow \mathrm{MN}(0, V), \text{ with } V = \frac{\omega_{\psi\psi.x}}{f\left(F^{-1}(\tau)\right)^{2} \int_0^1 B_x^2(r)\, dr}, \tag{2.22}$$

where $f(\cdot)$ is the density of u_{0t} which is assumed to be continuous, and $f\left(F^{-1}(\tau)\right)$ is the density evaluated at the τth quantile $F^{-1}(\tau)$ of that distribution. The quantity $\sigma_{\psi\psi.x} = \sigma_{\psi\psi} - \sigma_{\psi x}^2/\sigma_{xx}$ is the conditional variance of $\psi_{t\tau} = \psi_\tau(u_{0t\tau}) = \tau - \mathbf{1}\left\{u_{0t\tau} < F_{u_0}^{-1}(\tau)\right\}$, where $u_{0t\tau} = u_{0t} - F_{u_0,t-1}^{-1}(\tau)$ and

$$\mathbb{E}\left(\phi_t \phi_t' | \mathcal{F}_{t-1}\right) = \begin{bmatrix} \sigma_{\psi\psi} & \sigma_{\psi x} \\ \sigma_{x\psi} & \Sigma_{xx} \end{bmatrix}, \text{ with } \phi_t' = (\psi_{t\tau}, u_{xt}).$$

When $\beta(\tau) = \beta + \frac{b(\tau)}{d_n}$, we correspondingly have the following limit theory for the quantile FM estimated localizing coefficient $\hat{b}^{+}(\tau)$

$$\frac{n}{d_n}\left(\hat{b}^{+}(\tau) - b(\tau)\right) \Rightarrow \mathrm{MN}(0, V). \tag{2.23}$$

When $\frac{d_n}{n} + \frac{\sqrt{n}}{d_n} \to 0$ we then have specification validity in the sense that (2.21) holds asymptotically and FM regression asymptotics take the form (2.23). It is apparent that (2.22) can be used to construct pointwise confidence intervals for $\beta(\tau)$ for each value of τ. These may then be compared with estimates of β based on the null hypothesis that $\beta(\tau) = \beta$ is constant across quantiles.

As in the case of simple linear predictive regression (or cointegration), these nice asymptotics fail as soon as the time series x_t is LUR, in which case there are bias and nonstandard inference problems, just as in the linear predictive regression case. There are now new methods for addressing these difficulties which we discuss in the following section.

2.4 Misbalancing in Predictive Regressions

The linear predictive regression model (2.1) is a convenient formulation for practical work and is extensively used in applications with a large range of possible predictors that are often selected because of their plausibility as explanatory drivers of the dependent variable y_t. As such, the formulations used in practice are typically empirical with little attention given to their time series properties, a fact that can lead to problems of balance in the time series regression. For instance, if y_t has short memory and some of the regressors x_t have long memory, then a linear regression equation is potentially unbalanced. Many applications in finance are of this type. In the first place, these applications typically concentrate on predicting financial asset returns, which approximate martingale differences and are therefore hard to forecast, as attested in the two headers to the article. On the other hand, many of the potential predictors like interest rates have long memory or near unit root behavior which produce time series persistence characteristics in x_t that are very different from those of y_t. How such differences in the time series characteristics of the variables in a linear regression equation are reconciled is a major challenge in predictive regression research.

To fix ideas, suppose the predictive model formulation follows (2.1) and (2.2) where the regressor x_t is an LUR process with $\rho = 1 + \frac{c}{n}$ for some fixed c and u_{0t} is a martingale difference sequence (mds) with $E_{t-1}\left(u_{0t}^2\right) = \sigma_{00}$ *a.s.* . Under the null, $y_t = u_{0t}$ is also an mds. But under the alternative hypothesis of predictability where $\mathbb{H}_A : \beta = \beta_A \neq 0$, both y_t and $\beta_A x_{t-1}$ are $O_p\left(\sqrt{t}\right)$. So the equation implies different time series properties for y_t under the null and the alternative, meaning that the maintained formulation of the equation is unbalanced either under the null or the alternative, given a time series y_t with certain well defined characteristics. To illustrate, suppose y_t is an $I(0)$ series and x_t is $I(1)$ with $\rho = 1$ in (2.2). Then, (2.1) is unbalanced under the alternative with $\beta_A \neq 0$. Nonetheless, the fitted least squares coefficient is

$$\hat{\beta} = \frac{1}{n} \frac{\frac{1}{n}\sum_{t=1}^{n} x_{t-1} y_t}{\frac{1}{n^2}\sum_{t=1}^{n} x_{t-1}^2} = O_p\left(\frac{1}{n}\right) \to_p 0,$$

and the limit behavior $\hat{\beta} \to_p 0$ conflicts with the (true) alternative hypothesis $\mathbb{H}_A : \beta = \beta_A \neq 0$ and confirms the (false) null hypothesis $\mathbb{H}_0 : \beta = 0$ that accords with the observed $I(0)$ property of y_t. So, testing $\mathbb{H}_0$ against $\mathbb{H}_A$ reveals an explicit built-in contradition of the model because $\mathbb{H}_A$ is always incompatible with the observed $I(0)$ property of y_t. In forecasting stock returns, whose behavior is similar to an $I(0)$ series, primary interest lies usually in the alternative hypothesis $\mathbb{H}_A$, so that driver variables of returns can be identified. However, in view of the incompatibility of $\mathbb{H}_A$ with the observed $I(0)$ property of y_t, to accept the alternative effectively amounts to acceptance of a misspecified model.

Balancing issues such as the example above commonly arise in applied econometric work and sometimes originate in economic theory formulations or data identities. In the Fisher equation, for example, real and nominal interest rates are related linearly to expected inflation. In practical work, this relation involves the two latent variables, expected inflation and the (ex ante) real rate, which are frequently proxied by using the ex post real rate computed using the realized contemporaneous inflation rate. However, when we analyze the time series characteristics of these three observable variables (nominal rates, ex post real rates, and realized inflation) we frequently find major differences in the memory characteristics of the time series, even though the series are, by construction, linearly related. Such problems have been discussed in the literature - Phillips (2005) and Sun and Phillips (2004) - and they have no immediate or easy solution. In this example where the ex post real rate series is constructed directly from the nominal rate and realized inflation, the properties of the latter series are imposed on the former. So the ex post real rate must inherit, as a time series mixture of the other two series, at least some of their characteristics in terms of memory and heterogeneity. Nonetheless, when the memory characteristics of the individual series are estimated, there is no assurance (unless the requirement is imposed) that these memory characteristics will be compatible.

Similar issues arise in the context of predictive regression. One way of addressing these difficulties is to use localizing coefficients that assist in bringing the series into balance asymptotically. This asymptotic balancing can be achieved as follows. Suppose the predictive model 2.1() is replaced by

$$\begin{aligned} y_t &= \beta_n x_{t-1} + u_{0t}, \text{ with } \beta_n = \frac{b}{n^{\gamma}} \text{ for some } \gamma > 0 \qquad (2.24) \\ x_t &= \rho_n x_{t-1} + u_{xt}, \quad \text{with } \rho_n = 1 + \frac{c}{n} \text{ for some finite } c \in (-\infty, \infty). \end{aligned}$$

Then the localizing coefficient b measures marginal departues from the null. In this case we have the taxonomy

$$y_t = \begin{cases} u_{0t} + bn^{\frac{1}{2}-\gamma} J_x\left(\frac{t}{n}\right) = O_p\left(n^{\frac{1}{2}-\gamma}\right) & \gamma < 0.5 \\ u_{0t} + bJ_x\left(\frac{t}{n}\right) = O_p(1) & \gamma = 0.5 \\ u_{0t} = O_p(1) & \gamma > 0.5 \end{cases} . \tag{2.25}$$

These time series characteristics show that local departures from the null hypothesis $\mathbb{H}_0 : \beta = 0$ of the form $\mathbb{H}_A : \beta = \frac{b}{n^\gamma}$ are compatible with mds behavior of y_t asymptotically as long as the marginalizing rate coefficient $\gamma > 0.5$, or near mds behavior of y_t if $\gamma \geq 0.5$. For such near localizations to the null $\mathbb{H}_0$, the model (2.24) balances the $I(0)$ property of y_t with the $I(1)$ property of the regressor x_t.

Tests of $\mathbb{H}_0$ will then be consistent under the local alternative $\beta = \frac{b}{n^\gamma}$ provided $\gamma \in [0.5, 1)$ because of the divergent behavior of the quantity

$$n\hat{\beta}_n = n\left(\hat{\beta}_n - \beta_n\right) + n\beta_n = O_p(1) + O_p\left(n^{1-\gamma}\right), \tag{2.26}$$

where $\hat{\beta}_n$ is the least squares estimate of β_n in (2.24). On the other hand, local alternatives $\beta = \frac{b}{n^\gamma}$ with $\gamma > 1$ will be undetectible and tests of $\mathbb{H}_A$ will be inconsistent as they are too close to the null hypothesis in this case[2].

In sum, we find that under certain conditions marginal deviations from zero of the slope coefficient β_n in (2.24) are compatible with observed $I(0)$ or mds character in the dependent variable y_t and yet may still be distinguishable from the null hypothesis $\beta_n = 0$ in statistical testing. The key condition is that the marginal departures from the null must be small enough $\left(\beta_n = \frac{b}{n^\gamma} \text{ with } \gamma \geq 0.5\right)$ to retain the time series character of the observed y_t but not so small $(\gamma < 1)$ that they are indistinguishable from the null.

---

[2] The usual least squares t ratio is, under $\mathbb{H}_A : \beta_n = \frac{b}{n^\gamma}$,

$$t_\beta = \frac{\hat{\beta}_n}{s_{\hat{\beta}_n}} = \frac{n\hat{\beta}_n}{\left\{s^2\left(n^{-2}\sum_{t=1}^n x_{t-1}^2\right)\right\}^{1/2}} \sim \frac{n\hat{\beta}_n}{\left\{\sigma_{00}\left(n^{-2}\sum_{t=1}^n x_{t-1}^2\right)\right\}^{1/2}} = O_p\left(n^{1-\gamma}\right),$$

in view of (2.26) and since

$$s^2 = n^{-1}\sum_{t=1}^n \left(y_t - \hat{\beta}_n x_{t-1}\right)^2 = n^{-1}\sum_{t=1}^n u_{0t}^2 - n\left(\hat{\beta}_n - \beta_n\right)^2 n^{-2}\sum_{t=1}^n x_{t-1}^2 \to_p \sigma_{00},$$

as $\hat{\beta}_n - \beta_n = O_p\left(n^{-1}\right)$.

3 Recent Developments

We start our discussion by continuing with the linear predictive regression model (2.1) - (2.2). This framework is still the most popular in empirical work and is well suited to the new methodology of prediction using IVX endogenous instrumentation that we will consider first. The IVX approach applies to both short horizon and long horizon prediction, as well as in cases with multiple predictors. The second development involves nonparametric methods and is designed for a modeling framework that allows for more general functional forms in the predictive component.

3.1 IVX Endogenous Instrumentation

As explained earlier, one of the critical difficulties for inference and prediction in models such as (2.1) - (2.2) is the uncertainty that prevails in practical work about the degree of persistence in the predictor variables x_t. This uncertainty is commonly captured through the use of an LUR formulation and LUR asymptotics arising from an autoregressive coefficient specification $\rho = 1 + \frac{c}{n}$ with the unknown localizing coefficient c providing some modeling flexibility concerning the properties of x_t. Such a specification is convenient analytically but leads to the nonstandard inference complications discussed earlier. These become particularly troublesome in the multivariate predictor case which is important in practical work.

The new method of endogenous instrumentation is designed to address these difficulties and was suggested in Phillips and Magdalinos (2009). The idea is to use the (endogenous) regressors x_t to self-generate instrumental variables (hence, the terminology IVX) with properties that remove the parameter dependencies and distributional complexities of LUR asymptotics. The intention is to bypass these difficulties by creating instruments from x_t that have less persistence and, more especially, less persistence than regressors with a UR or LUR form. The cost of reducing persistence in the case of UR and LUR predictors is reduction in the convergence rate of the estimator from the usual $O(n)$ rate. The IVX instruments are simple to construct using an autoregressive recursion.

As before it is convenient to illustrate the workings with the scalar predictor case, although the method applies equally well with no further computation in the multivariate case. The self-generated instruments are obtained by differencing the predictor x_t and using the following autoregressive filter to construct the (mildly integrated) instruments

$$\tilde{z}_t = \sum_{j=1}^{t} \rho_{nz}^{t-j} \triangle x_j, \quad \text{with } \rho_{nz} = 1 + \frac{c_z}{n^{\varphi}}, \ \varphi \in (0,1), \ c_z < 0. \tag{3.1}$$

The autoregressive coefficient ρ_{nz} in this filter is selected to lie in the mildly integrated zone (Phillips and Magdalinos, 2007), since $c_z < 0$ and $\varphi < 1$, thereby ensuring that the IVX instrument is mildly integrated and less persistent than a UR or LUR regressor. In fact, the IVX instrument $\tilde{z}_t$ may be used whether x_t has a unit root, local unit root, or is itself mildly integrated or mildly explosive. In this sense, IVX instruments like $\tilde{z}_t$ offer considerable robustness because the inferential procedure relies only on standard asymptotics and completely avoids problems of nonstandard inference. In what follows, it will be convenient to assume that x_t is a LUR predictor. This case, as well as UR and mildly integrated cases, are treated in full in Phillips and Magdalinos (2009) and Kostakis et al (2015).

When x_t is LUR, $\triangle x_t = \frac{c}{n} x_{t-1} + u_{xt}$ and $\tilde{z}_t$ decomposes as

$$\tilde{z}_t = \sum_{j=1}^{t} \rho_{nz}^{t-j} u_{xt} + \frac{c}{n} \sum_{j=1}^{t} \rho_{nz}^{t-j} x_{j-1} =: z_t + \frac{c}{n} \psi_{nt}, \tag{3.2}$$

from which it is apparent that $z_t = \rho_{nz} z_{t-1} + u_{xt}$ plays the role of a mildly integrated instrument that is approximated in practical implementation by $\tilde{z}_t$. The approximation holds because the remainder term $\frac{c}{n}\psi_{nt}$ in (3.2) turns out to be negligible in all cases other than when x_t is mildly explosive in which case the IVX instrument $\tilde{z}_t$ is still effective in inference, as shown in recent work (Phillips and Lee, 2015b).

Using $\tilde{z}_{t-1}$ as an instrument for x_{t-1} in (2.1) leads by means of the usual IV regression formula to the estimate

$$\hat{\beta}_{IVX} = \frac{\sum_{t=1}^{n} \tilde{z}_{t-1} y_t}{\sum_{t=1}^{n} \tilde{z}_{t-1} x_{t-1}} = \beta + \frac{\sum_{t=1}^{n} \tilde{z}_{t-1} u_{0t}}{\sum_{t=1}^{n} \tilde{z}_{t-1} x_{t-1}}. \tag{3.3}$$

The IVX estimator (3.3) is particularly simple in the present case because the equation error is an mds.[3] The estimation error involves the sample covariance $\sum_{t=1}^{n} \tilde{z}_{t-1} u_{0t}$ between

---

[3] Otherwise, a one-sided long run covariance correction term, just as in FM regression (Phillips and Hansen, 1990), is introduced to deal with serial correlation induced by weakly dependent errors. For details, see Phillips and Magdalinos (2009).

the equation error u_{0t} and the IVX instrument $\tilde{z}_{t-1}$. The explanation for the simplicity of IVX asymptotics is that the troublesome (nonstandard) second term of the usual LUR limit theory (2.8) is eliminated. More specifically, with an LUR predictor and AR coefficient $\rho = 1+\frac{c}{n}$, the limiting sample covariance $\frac{1}{n}\sum_{t=1}^{n} x_{t-1}u_{0t}$ converges weakly to the stochastic integral $\int_0^1 J_x^c(s)dB_0(s)$ whose correlated stochastic processes J_x^c and B_0 lead to the troublesome nonstandard component in (2.8) when $\sigma_{0x} \neq 0$. In contrast, after normalization the sample IVX covariance $\sum_{t=1}^{n} \tilde{z}_{t-1}u_{0t}$ satisfies a martingale central limit theorem and is asymptotically independent of the IVX signal (relevance) quantity $\sum_{t=1}^{n} \tilde{z}_{t-1}x_t$. Using this result, Phillips and Magdalinos (2009) show under some regularity conditions that

$$n^{(1+\varphi)/2}\left(\hat{\beta}_{IVX} - \beta\right) \Rightarrow MN\left(0, \sigma_{00}\Psi\right), \tag{3.4}$$

where the mixed normal (MN) limit distribution enables pivotal inference of the predictability hypothesis $\mathbb{H}_0 : \beta = 0$ using standard t and Wald test statistics even when the limit quantity Ψ involves random elements, as it does when $\rho = 1 + \frac{c}{n}$.

The mixed normal asymptotics of $\hat{\beta}_{IVX}$ in (3.4) are correctly centred at β, so there is no asymptotic bias. The convergence rate $O\left(n^{\frac{1+\varphi}{2}}\right)$ of $\hat{\beta}_{IVX}$ is less than $O(n)$ and depends on the localizing power parameter $\varphi < 1$ that is used in self-generating the instruments $\tilde{z}_t$. The (random) variance quantity Ψ in (3.4) is estimated via the IVX signal $\sum_{t=1}^{n} \tilde{z}_{t-1}x_t$ and σ_{00} is estimated from the regression residuals in the usual way. Testing is then conducted by means of t ratios (or Wald tests in the multivariate case) with convenient standard normal (or chi square) limit theory. In the present scalar case, we have, quite simply,

$$t_{\hat{\beta}_{IVX}} = \frac{\hat{\beta}_{IVX} - \beta}{\widehat{\sigma}_{IVX}} \Longrightarrow \xi =_d N\left(0, 1\right), \tag{3.5}$$

where $\widehat{\sigma}_{IVX}$ is the standard error of $\hat{\beta}_{IVX}$ computed via the conventional formula $\widehat{\sigma}_{IVX}^2 = \widehat{\sigma}^2\left(\sum_{t=1}^{n} \tilde{z}_{t-1}^2\right)\left(\sum_{t=1}^{n} \tilde{z}_{t-1}x_{t-1}\right)^{-2}$, with $\widehat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} \hat{u}_{0t}^2$ based on the IVX residuals $\hat{u}_{0t} = y_t - \hat{\beta}_{IVX}x_t$. When the errors u_{0t} are conditionally heteroskedastic, the usual correction to $\widehat{\sigma}_{IVX}^2$ can be employed to ensure the validity of (3.5), viz.,

$$\tilde{\sigma}_{IVX}^2 = \left(\frac{1}{n}\sum_{t=1}^{n} \tilde{z}_{t-1}^2 \hat{u}_{0t}^2\right)\left(\sum_{t=1}^{n} \tilde{z}_{t-1}x_{t-1}\right)^{-2}.$$

Implementation of the IVX estimator (3.3) and the test (3.5) requires use of the filter

(3.1), which in turn requires specification of the parameter ρ_{nz} and hence the localizing exponent φ. The localizing coefficient c_z in ρ_{nz} can be set to $c_z = -1$, so that the degree of mild integration is controlled entirely by φ. When x_t is a UR or LUR process, the limit theory (3.5) holds for all $\varphi \in (0, 1)$ and the convergence rate of $\hat{\beta}_{IVX}$ is $O\left(n^{\frac{1+\varphi}{2}}\right)$, so selection of φ close to unity is preferable bringing the rate close to the optimal rate $O\left(n^{-1}\right)$. In finite samples, power tends to increase monotonically with φ and size is very well controlled until φ is very close to unity. After extensive simulation experiments, Kostakis et al (2015) recommend the choice $\varphi = 0.95$ to ensure size is well controlled and power is close to maximal. Analytic methods for the optimal choice of φ and data-based algorithms for its selection are desirable, but presently unavailable. New initiatives are needed to find such rules of selection, as the usual methods of optimizing asymptotic mean square error are known to fail in this case (Phillips and Lee, 2015b).

Simulations show that inference on predictability using (3.5) and its corresponding Wald statistic extensions in the multivariate case all work well in practice with good size and power properties for predictors in the UR, LUR, and mildly integrated range. Kostakis et al (2015) and Phillips and Chen (2014) report some extensive Monte Carlo experiments investigating the performance of IVX in comparison with other procedures. In particular, comparisons with the Campbell and Yogo (2006) and Jansson and Moreira (2006) methods that were discussed above indicate that IVX inference has better size, accommodates a much wider range of possible predictors, extends easily to multivariate settings where those methods are unavailable, and generally has superior power properties. The method is easily implemented in the case where an intercept is fitted in (2.1) in which case a minor but important modification to the test can be made to improve finite sample size performance (Kostakis et al, 2015). Further recent work (Phillips and Lee, 2015b) shows the IVX tests remain valid in cases where there are mixed orders of persistence in the predictors.

The IVX approach also extends readily to long horizon prediction, where interest centres on predictions more than one period ahead and often on far horizon predictions K periods ahead. Analytic work on the use of IVX methods in such cases has been done in Phillips and Lee (2013) and Kostakis et al (2015) using slightly different approaches. Phillips and Lee work, as in much of the literature on long horizon predictions, with a temporally aggregated version of the model (2.1) as well as temporally aggregated IVX instruments. Kostakis et al work with the temporally aggregated model but retain the usual IVX instruments. Both methods produce standard asymptotics for inference, analogous to (3.5) for t tests and chi-square for Wald tests of predictability. Importantly, these methods are also robust to far horizon cases modeled as $K \to \infty$ at a rate slower than $n \to \infty$.

IVX methods have been used with success in recent applied work, particularly in research on stock market predictability where it is useful to be able to assess the statistical importance of several posited predictors. As the headers to the paper emphasize, investigators are especially interested in knowing not only whether excess returns in the stock market are predictable, but also which of the many financial indicators now available are useful in delivering good predictions. The flexibility and ready implementability of IVX methods make them attractive in such exercises. In their extensive empirical application allowing for single and multiple potential predictors of US stock returns, Kostakis et al (2015) interestingly find less evidence for long horizon than short horizon predictability, concluding that

> "our long-horizon tests document that, if anything, predictability becomes weaker, not stronger, as the horizon increases"

This conclusion supports some of the early concerns about doubtful evidence of stock market forecasting capability raised by Alfred Cowles in the primary header to this article, concerns which were recently seconded in the study by Welch and Goyal (2007), and that stands in contrast to the affirmation of long run market predictability given in the second header by the Royal Swedish Academy.

Closely related methods to IVX that use modifed variable addition (VA) regressions have most recently been introduced by Breitung and Demetrescu (2015) where lagged predictors are replaced by persistent time series using the Phillips-Magdalinos methodology of self-generated variables and instrumentation. These VA methods, like IVX, help remove non-pivotal inference problems when there are LUR predictors, and are similarly applicable when there are multiple predictors.

3.2 Nonparametric Predictive Regression

All of the methods so far discussed are parametric and involve linear model specifications. For predictive regression modeling, just as for other areas of applied econometric work, linear relationships may be convenient in practice but may only provide a first approximation to a nonlinear behavioral response. In financial market prediction, we may well expect such responses to entail nonlinearities, if only because of differences in response to positive and negative financial indicators. Moreover, as noted earlier, linear model specifications are typically unbalanced in terms of the respective memory properties of the time series and therefore require localization of the coefficients to restore long term balance. Such localizations may be regarded as a form of nonlinearity in which the coefficients take different values according to the sample size to ensure plausible properties in the predictive equation. In cases like these where nonlinearities in behavioral response do occur or are anticipated, it seems appropriate to use methods that accommodate potential nonlinearities directly. Further, test size based on linear predictive model fitting seems unlikely to be robust to functional form misspecification and test power may be quite sensitive to functional form.

Some of these issues have been investigated in recent work by Kasparis, Andreou and Phillips (2014; KAP), who propose a unifying framework for predictive inference that allows for the possibility of nonlinear relationships of unknown form. The prediction tests suggested in this work rely on nonparametric kernel estimation methods and offer robustness to integration order, including fractional orders, as well as functional form. The methods draw on and develop in certain respects other recent econometric work on nonparametric kernel regression with nonstationary time series.

In place of (2.1) we consider the nonlinear predictive system

$$\begin{aligned} y_t &= g\left(x_{t-1}\right) + u_{0t}, && (3.6) \\ x_t &= \left(1+\frac{c}{n}\right) x_{t-1} + u_{xt}, && (3.7) \end{aligned}$$

where $g\left(\cdot\right)$ is some smooth unknown regression function that provides the systematic predictive response of y_t to the past history of the predictor x_t embodied in its past value x_{t-1}, and with initialization $x_0 = O_p\left(1\right)$. When x_t is a stationary weakly dependent process, rather than the LUR process given in (3.7) the limit theory of nonparametric regression estimators for models such as (3.6) follows from standard theory and pivotal testing of nonlinear prediction follows directly in such cases. It is now known from Wang and Phillips (2009, 2015) that, somewhat remarkably, this standard limit theory for kernel estimation continues to apply in cases where the predictor follows a UR or LUR time series as in (3.7) and is an endogenous regressor, although with reduced rates of convergence. KAP use this theory and some extensions of it involving long memory and antipersistent innovations in (3.7), to show that predictability tests can be mounted using kernel estimates of (3.6). These tests have standard asymptotics, are robust across a variety of generating mechanisms for the predictor variable, and are easily implemented in practical work.

An important advantage of the specification (3.6) is that certain nonlinear transformations of nonstationary time series such as the LUR process (3.7) produce new trajectories that have a character closer to a stationary series than an LUR time series, thereby attenuating issues of balance in the specification of the predictive regression under the alternative. For example, integrable transforms of a unit root process are known to substantial reduce the persistence properties of the original series, producing a new series with memory parameter $d \sim \frac{1}{4}$, which lies in the stationary zone (see, e.g., Marmer, 2007; Miller and Park, 2010; Kasparis, Phillips and Magdalinos, 2014). In such cases, the output series y_t can be stationary, even when the input series x_{t-1} has considerable time series persistence. To illustrate, Figure 2 displays 500 observations of a standard Gaussian random walk time series x_t, its integrable exponential transform $g\left(x_t\right) = \exp\left(-\frac{1}{2}x_t^2\right)$, and the same transform with additive white noise $u_{0t} \sim_{iid} N\left(0,1\right)$. Apparently, the transformed series considerably reduces the random wandering behavior of x_t, attenuating its signal, and producing a new series that is much closer to the origin with departures occuring primarily in those regions where the random walk x_t is in the vicinity of the origin. The transformed series with additive noise appears like a stationary time series centred on the origin with some tendency occasionally to drift away from the origin, much like that of a stationary long memory series.

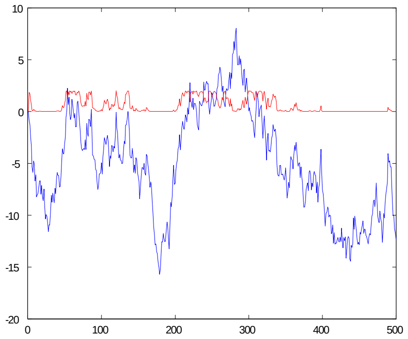

Figure 2a: Simulated trajectories of a random walk x_t (blue) and its exponential integrable transform $\exp\left(-\frac{1}{2}x_t^2\right)$ (red).

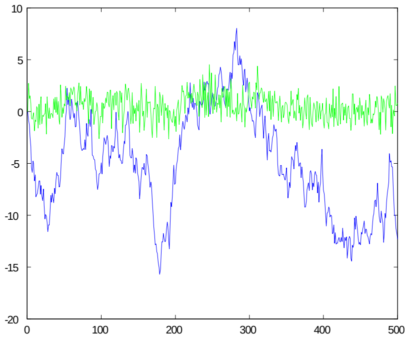

Figure 2b: Simulated trajectories of a random walk x_t (blue) and its transform with additive noise $\exp\left(-\frac{1}{2}x_t^2\right) + u_{0t}$ (green)

KAP consider a model of the form (3.6) - (3.7) in which the error u_{xt} in (3.7) may be a short-memory (SM) time series or a stationary $ARFIMA(d)$ time series with memory parameter $d \in (-1/2, 1/2)$, allowing for either long memory (LM) or anti-persistence (AP). The regression function g in (3.6) is estimated by simple kernel regression giving

$$\hat{g}(x) = \frac{\sum_{t=1}^{n} K_h\left(x_{t-1} - x\right) y_t}{\sum_{t=1}^{n} K_h\left(x_{t-1} - x\right)}, \tag{3.8}$$

where $K_h(\cdot) = \frac{1}{h}K(\cdot/h)$, $K(\cdot)$ is a suitable kernel function, and h is a bandwidth parameter for which $h = h_n \to 0$ as $n \to \infty$. Under conditions based on those of Wang and Phillips (2009), KAP show that the nonparametric estimator $\hat{g}(x)$ has the following self-normalized limit theory as $n \to \infty$

$$\left(\sum_{t=1}^{n} K\left(\frac{x_{t-1} - x}{h_n}\right)\right)^{1/2} \left(\hat{g}(x) - g(x)\right) \xrightarrow{d} N\left(0, \sigma_{00} \int_{-\infty}^{\infty} K(s)^2 ds\right). \tag{3.9}$$

Hence, in the predictive regression framework (3.6)-(3.7), $\hat{g}(x)$ is consistent and has a Gaussian limit distribution that is free of the nuisance parameter c in the LUR specification (3.7). This limit theory is the same as when x_t is a stationary weakly dependent process such as a stable AR process. Thus, (3.9) offers wide generality in the predictive regression context, allowing for many predictor processes x_t that include LUR and LM time series, which facilitates the development of a class of nonparametric predictability tests.

Under the null hypothesis of no predictability in (3.6) the regression function is a constant, so that $y_t = \mu + u_{0t}$ giving the null formulation $\mathbb{H}_0 : g(x) = \mu$. Hence, in view of (3.9), $\hat{g}(x) \to_p \mu$, which suggests a test based on

$$\hat{t}(x, \mu) := \left(\frac{\sum_{t=1+\nu}^{n} K\left(\frac{x_{t-\nu} - x}{h_n}\right)}{\hat{\sigma}_{00} \int_{-\infty}^{\infty} K(s)^2 ds}\right)^{1/2} \left(\hat{g}(x) - \mu\right), \tag{3.10}$$

where $\hat{\sigma}_{00} = n^{-1} \sum_{t=1}^{n} \left(y_t - \hat{\mu}\right)^2$ is a consistent estimator of σ_{00}. The idea is to compare the estimator $\hat{g}(x)$ with a constant function and, although μ is generally unknown, it can be consistently estimated at a $\sqrt{n}$ rate under the null by the sample mean $\hat{\mu} = n^{-1} \sum_{t=1}^{n} y_t$ which ensures that $\hat{t}(x, \hat{\mu}) = \hat{t}(x, \mu) + o_p(1)$ and leads to the following straightforward limit theory

$$\hat{t}(x, \hat{\mu}) \xrightarrow{d} N\left(0, 1\right), \tag{3.11}$$

which compares the nonparametric estimate $\hat{g}(x)$ with the parametric estimate $\hat{\mu}$.

Predictive test statistics are based on making the comparison in (3.11) over some point set to assess constancy of the predictive function over this set. In particular, let X_s be a set of isolated points $X_s = \{\bar{x}_1, ..., \bar{x}_s\}$ in $\mathbb{R}$ for some fixed $s \in \mathbb{N}$. The tests proposed in KAP involve sum and sup functionals over this set, viz.,

$$\widehat{F}_{\text{sum}} := \sum_{x \in X_s} \hat{F}(x, \hat{\mu}) \text{ and } \widehat{F}_{\max} := \max_{x \in X_s} \hat{F}(x, \hat{\mu}), \text{ with } \hat{F}(x, \hat{\mu}) := \hat{t}(x, \hat{\mu})^2. \tag{3.12}$$

In practical work the set X_s can be chosen using uniform draws over some region of particular interest in the state space. The null distributions of these test statistics follow from (3.11)

$$\widehat{F}_{\text{sum}} \rightarrow_d \chi_s^2 \textit{ and } \widehat{F}_{\max} \rightarrow_d Y,$$

where the random variable Y has c.d.f. $F_Y(y) = P(X \leq y)^s$ with $X \sim_d \chi_1^2$. Thus, the limit distributions of the tests involve functionals of independent chi squared variates which are readily computed. KAP show that this limit theory holds for a wide class of predictors x_t that includes LUR and LM time series, thereby allowing for an extensive range of persistent regressors in (3.6).

This approach to predictive regression using nonparametric kernel regression and grid testing for constancy in the regression has appeal in terms of its robustness to the generating mechanism of the predictor. The framework helps to unify predictive inference in situations where both model functional form and the properties of the predictor are unknown. Simulations reported in KAP show stable size performance for both tests and good power in comparison with other procedures even against linear alternatives. The nonparametric tests are decidedly superior against nonlinear predictive model alternatives, as might be expected, and perform well for both long memory and LUR predictors. One disadvantage of these tests is that they are mainly useful in cases where the predictor is a scalar time series, like many of the procedures in current use. Interestingly, the nonparametric approach which provides these tests with their generality over such a wide range of predictor processes, also typically delimits applications to predictive regressions with a single regressor or to additive nonlinear functionals because of the curse of dimensionality.

Nonparametric techniques deliver smooth predictor functions for arbitary x by virtue of the kernel estimated form $\hat{g}(x) = \sum_{t=1}^n K\left(\frac{x_{t-1}-x}{h_n}\right) y_t / \sum_{t=1}^n K\left(\frac{x_{t-1}-x}{h_n}\right)$, which has the same smoothness properties as the kernel function K. When the alternative hypothesis holds and there is predictability from $g(x)$, we might expect the estimate $\hat{g}(x)$ to be a better predictor within, rather than outside, the sample space. But when x_t is recurrent, the sample space is inevitably large and integrable nonlinear functions g attenuate the effects of outlier realizations of x_t. So the impact of outliers on g is controlled by functional form, as required by the balancing of the regression function. This attentuation continues to apply in prediction. Thus, in the case of financial asset return prediction by persistent regressors, the predictive capability of $g\left(x_t\right)$ tends to increase when x_t takes realizations around the origin.

More detailed information about the bias characteristics of the predictor $\hat{g}(x)$ can be deduced from nonparametric regression asymptotics, as we now briefly discuss. Wang and Phillips (2015) give the following bias corrected form of (3.9)

$$\left(\sum_{s=1}^{n} K(\frac{x_{s-1}-x}{h_n})\right)^{1/2}\left[\hat{g}(x)-g\left(x\right)-\frac{h_n^2 g''(x)\mu_{2K}}{2}\right]\Rightarrow N\left(0,\sigma_{00}\int_{-\infty}^{\infty}K({}^2y)dy\right),$$

where $\mu_{2K}=\int_{-\infty}^{\infty}y^2K(y)dy$. The prediction bias of $\hat{y}_{n+1}=\hat{g}(x_n)$ is therefore $O\left(h_n^2\right)$ and the prediction error variance, conditional on $x=x_n$, is $O_p\left(\left(\sum_{s=1}^{n}K(\frac{x_{s-1}-x_n}{h_n})\right)^{-1}\right)$. To find the order of magnitude of this prediction error variance, assume the process x_t satisfies the functional law $n^{-1/2}x_{\lfloor n\cdot\rfloor}\Rightarrow G\left(\cdot\right)$ for some Gaussian stochastic process G whose local time $\ell_G\left(t,a\right)$ at the spatial point a over the time interval $[0,t]$ is given by (c.f. Revuz and Yor, 1999)

$$\ell_G\left(t,a\right)=\lim_{\epsilon\to0}\frac{1}{2\epsilon}\int_0^t\mathbf{1}[|G(s)-a|<\epsilon]ds.$$

Then, $\left(nh_n^2\right)^{-1/2}\sum_{s=1}^{n}K(\frac{x_{s-1}-x_n}{h_n})\Rightarrow\Sigma_{xx}^{-1/2}\ell_G\left(1,G\left(1\right)\right),$ where $\ell_G\left(1,G\left(1\right)\right)$ is the local time that the Gaussian process $G\left(s\right)$ has spent over the time interval $[0,1]$ at its final position $G\left(1\right)$ – see Phillips (2009) and Wang and Phillips (2009, 2012). In this event, the prediction error variance of $\hat{y}_{n+1}=\hat{g}(x_n)$ has the order of magnitude $O_p\left(\left(\sum_{s=1}^{n}K(\frac{x_{s-1}-x_n}{h_n})\right)^{-1}\right)=O_p\left(\left(nh_n^2\right)^{-1/2}\right)$, a rate which reflects the slow convergence rate of the nonparametric estimate $\hat{g}$.

4 Conclusion

> "*The plain truth is that facts are only facts; for predicting the effects of economic changes they cannot take the place of relationships between economic variables.*" Johnson (1960)

> *"General economics laws are unhelpful as a guide to understand the past or predict the future because they ignore the political and economic institutions, as well as the endogenous evolution of technology, in shaping the distribution of resources in society"* Acemoglu and Robinson (2015)

Whatever the pitfalls and difficulties that have been discovered with predictive regressions and analyzed herein, quantitative assessments of predictability using modern econometric methods of bias reduction, endogenized instrumentation, and quantile regression have, nonetheless, sound inferential underpinnings. This basis enhances confidence in the use of the methods in practical work, if only about the uncertainty of the predictions. A firm statistical foundation is especially useful in determining drivers of economic time series such as asset price returns where the effects of fundamentals are so frequently obscured in short run volatility and where there are so many potential determinants that vie for inclusion. In comparison, the alternative approach of relying completely on descriptive and qualitative appraisals of significance is generally unhelpful in transfering knowledge and in defining the region of uncertainty or ignorance about the phenomenon under study and the predictions being made. As Johnson (1960) aptly described the matter more than half a century ago in the headnote of this conclusion, relationships between economic variables inevitably play the critical role in making scientific predictions. Data alone is insufficient, even in the context of financial markets where its prodigal abundance has raised quite new 'big data' and degrees of freedom (or so-called $p > n$) problems of statistical modeling and inference.

Outside of financial applications, predictive regressions play a significant role in diverse areas of applied econometric work. Many of the big questions being addressed at present in the macroeconomic arena, for instance, involve trending economic variables, such as the patterns, drivers and predictors of economic growth, issues of growth convergence, and the relationship of growth to the evolving nature of inequality in both income and wealth inequality. In this field, Acemoglu and Robinson (2015) describe the deep institutional complexities that underly some of these economic issues in their essay on the rise and decline of capitalism. They cite the difficulties in predicting the future because of the politico-economic-institutional complexity of modern society and inherent endogeneities in the technology that underlies production and income generation. The upshot is this: however much as econometricians we wish to follow Johnson's dictum about utilizing relationships among economic variables in order to make predictions, the task is bedeviled by the complexities, interdependencies, and evolutionary nature of the data we have available to use. In short, successful predictive regression, just like econometric model-building in general, inevitably steers a harrowing course between, as Cragg (1968) aptly once described it, "the Scylla of specification error and the Charybdis of underspecification".

Throughout his career as an econometrician, Halbert White was absorbed with the task of developing econometric methods of estimation and inference that are robust to misspecification. Following the impetus of his work, this line of research blossomed and has now infiltrated virtually every arena of econometric work, including forecasting. Indeed, many of the techniques discussed in this paper were influenced by the same concerns that motivated Halbert White's research and John Cragg's early warnings in the 1960s about specification error and underspecification in empirical econometric work.

5 References

Acemoglu, D. and J. A. Robinson, 2015. "The Rise and Decline of General Laws of Capitalism," *Journal of Economic Perspectives*, 29, 3-28.

Amihud, Y., and C. Hurvich (2004). "Predictive regressions: a reduced-bias estimation method," *Journal of Financial and Quantitative Analysis,* 39, 813–841.

Bak, P., C. Tang, and K. Wiesenfeld (1987). "Self-organized criticality: an explanation of 1/f noise". *Physical Review Letters,* 59, 381–384.

Breitung, J., and M. Demetrescu (2015). Instrumental variable and variable addition based inference in predictive regressions. *Journal of Econometrics,* 187(1), 358-375.

Campbell, J. and M. Yogo (2006). "Efficient tests of stock return predictability," *Journal of Financial Economics*, 81(1), 27-60.

Cavanagh, C., G. Elliott. and J.Stock (1995). "Inference in models with nearly integrated regressors," *Econometric Theory*, 11(05), 1131-1147.

Cenesizoglu, T and A. Timmermann (2008). "Is the distribution of stock returns predictable?" Unpublished Manuscript, HEC Montreal and UCSD.

Chan, N.H. and C. Z. Wei (1987). "Asymptotic inference for nearly nonstationary AR(1) processes". *Annals of Statistics,* 15, 1050–1063.

Chen, W. W., & Deo, R. S. (2009). Bias reduction and likelihood-based almost exactly sized hypothesis testing in predictive regressions using the restricted likelihood. *Econometric Theory*, 25(05), 1143-1179.

Cowles, A. (1933), "Can stock market forecasters forecast?" *Econometrica*, 1, 309-324.

Cowles, A. (1944), "Stock market forecasting," *Econometrica*, 12, 206-214.

Cragg, J.G. (1968). Some effects of incorrect specification on the small-sample properties of several simultaneous-equation estimators. *International Economic Review*, 9, 63-85.

Elliott, G. and J.H. Stock (1994). "Inference in time series regression when the order of integration of a regressor is unknown," *Econometric Theory*, 10(3-4), 672-700.

Hahn, J. and G. Kuersteiner (2002) "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both N and T are large", *Econometrica*, 70, 1639–1657.

Han, C. and P. C. B. Phillips (2013). "First Difference MLE and Dynamic Panel Estimation", *Journal of Econometrics*, 175, 35–45.

Han, C., P. C. B. Phillips and D. Sul, (2014). "X-Differencing and Dynamic Panel Model Estimation," *Econometric Theory*, 30, pp 201-251.

Hansen, B. E. (1995): "Rethinking the univariate approach to unit root tests: How to use covariates to increase power," *Econometric Theory*, 11, 1148-1171.

Jansson, M. and M. Moreira (2006). "Optimal inference in regression models with nearly integrated regressors," *Econometrica*, 74, 681-714.

Jeganathan, P. (1995). "Some Aspects of Asymptotic Theory with Applications to Time Series Models," *Econometric Theory*, 11, 818–887.

Jeganathan, P. (1997). "On Asymptotic Inference in Linear Cointegrated Time Series Systems," *Econometric Theory*, 13, 692–745.

Kasparis, I., P. C. B. Phillips and T. Magdalinos (2014) "Nonlinearity Induced Weak Instrumentation" *Econometric Reviews*, 33, 676-712.

Kasparis, I., E. Andreou, and P. C. B. Phillips (2015). “Nonparametric Predictive Regression,” *Journal of Econometrics*, 185, 468-494.

Koenker, R and G. Basset (1978). “Regression quantiles,” *Econometrica,* 46, 33-49.

Kostakis, A., A. Magdalinos and M. Stamatogiannis (2015). “Robust econometric inference for stock return predictability," *Review of Financial Studies* (forthcoming).

Kothari, S.P., and J. Shanken (1997). "Book-to-market, dividend yield, and expected market returns: A time series analysis," *Journal of Financial Economics*, 18, 169-203.

Lamperti, J. (1958). “An occupation time theorem for a class of stochastic processes,” Transactions of the American Mathematical Society, 88, 380–387.

Lee, J.H. (2015). "Predictive Quantile Regression with Persistent Covariates: *IVX-QR Approach*," Journal of Econometrics (forthcoming)

Magdalinos, T. and P. C. B Phillips (2009). “Limit theory for cointegrated systems with moderately integrated and moderately explosive regressors”. *Econometric Theory*, 25, 482-526.

Marmer, V., (2007). Nonlinearity, nonstationarity and spurious forecasts. *Journal of Econometrics*, 142. 1-27.

Maynard A., K. Shimotsu and Y. Wang (2011). “Inference in predictive quantile regressions,” Unpublished Manuscript.

Miller, J. I., Park, J. Y. (2010). “Nonlinearity, nonstationarity, and thick tails: How they interact to generate persistence in memory,” J*ournal of Econometrics,* 155:83–89.

Moon, H. R. and P. C. B. Philllips (2000): “Estimation of Autoregressive Roots Near Unity Using Panel Data,” *Econometric Theory*, 16, 927–997.

Moon, H. R. and P. C. B. Phillips (2004): “GMM Estimation of Autoregressive Roots Near Unity with Panel Data,” *Econometrica*, 72, 467–522.

Park, J. Y. (2002). “Nonlinear nonstationary heteroskedasticity,” *Journal of Econometrics* 110:383–415.

Park, J.Y. and P.C.B. Phillips (1999). "Asymptotics for nonlinear transformations of integrated time series," *Econometric Theory*, **15**, 269-298.

Park, J.Y. and P.C.B. Phillips (2000). "Nonstationary binary choice," *Econometrica*, 68, 1249-1280.

Park, J.Y. and P.C.B. Phillips (2001). "Nonlinear regressions with integrated time series," *Econometrica,* **69,** 117-161.

Phillips, P. C. B. (1987a). "Time Series Regression with a Unit Root," *Econometrica,* 55, 277–302.

Phillips, P. C. B. (1987b). "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika* 74, 535–547.

Phillips, P. C. B. (1989). "Partially identified econometric models," *Econometric Theory* 5, 181–240.

Phillips, P. C. B. (1991). "Optimal Inference in Cointegrated Systems," *Econometrica* 59, 283–306.

Phillips, P. C. B. (2005). "Econometric analysis of Fisher's equation," *The American Journal of Economics and Sociology*, 64, 125 - 168.

Phillips, P. C. B. (2009). "Local limit theory and spurious nonparametric regression," *Econometric Theory*, 25 1466–1497.

Phillips, P. C. B. (2014). "On confidence intervals for autoregressive roots and predictive regression", *Econometrica*, 82, 1177-1195.

Phillips, P. C. B. and B. E. Hansen (1990). "Statistical inference in instrumental variables regression with I(1) processes," *Review of Economic Studies* 57, 99–125.

Phillips, P. C. B. and J-H. Lee (2013). "Predictive Regression under Varying Degrees of Persistence and Robust Long-Horizon Regression", *Journal of Econometrics*, 177, 250-264.

Phillips, P. C. B. and J.H. Lee (2015a). "Limit Theory for VARs with Mixed Roots Near Unity," *Econometric Review*s, 34, 1035-1056.

Phillips, P. C. B. and J-H. Lee (2015b). “Robust Econometric Inference with Mixed Integrated and Mildly Explosive Regressors”, Cowles Foundation Discussion Paper (revised), Yale University.

Phillips, P. C. B. and T. Magdalinos (2007a), "Limit theory for moderate deviations from a unit root," *Journal of Econometrics* 136, 115-130.

Phillips, P. C. B. and T. Magdalinos (2007b), "Limit theory for moderate deviations from a unit root under weak dependence," in G. D. A. Phillips and E. Tzavalis (Eds.) *The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis.* Cambridge: Cambridge University Press, pp.123-162.

Phillips, P. C .B., and T. Magdalinos, (2009). Econometric inference in the vicinity of unity.” CoFie Working Paper No. 7, Singapore Management University.

Phillips, P. C. B., S. Shi and J. Yu (2014). “Specification sensitivity in right-tailed unit root testing for explosive behaviour,” *Oxford Bulletin of Economics and Statistics*, 76, 315-333.

Phillips, P. C. B. and V. Solo (1992). “Asymptotics for linear processes”, *The Annals of Statistics*, pp. 971-1001.

Piketty, T. (2014). *Capital in the Twenty-First Century.* Harvard University Press.

Revuz, D. And M. Yor (1999). *Continuous Martingales and Brownian Motion.* New York: Springer–Verlag.

Stambaugh, R. (1999) “Predictive regressions,” *Journal of Financial Economics,* 54, 375–421.

Stock, J. (1991). "Confidence intervals for the largest autoregressive root in US macroeconomic time series," *Journal of Monetary Economics*, 28(3), 435-459.

Sun, Y., and P. C. B. Phillips (2005). "Understanding the Fisher Equation," *Journal of Applied Econometrics* 19, 2004, 869-886.

Torous, W. and R. Valkanov (2000)." Boundaries of predictability: Noisy predictive regressions," Unpublished Manuscript, UCLA.

Valkanov, R. (2003)."Long-horizon regressions: Theoretical results and applications," *Journal of Financial Economics,* 68(2), 201-232.

Wang, Q. and P.C.B. Phillips (2009). Structural nonparametric cointegrating regression. *Econometrica,* 77, 1901-1948.

Wang, Q. and P. C. B. Phillips (2015). "Nonparametric Cointegrating Regression with Endogeneity and Long Memory", *Econometric Theory* (forthcoming)

Welch, I., and A. Goyal, (2008)."A comprehensive look at the empirical performance of equity premium prediction," *Review of Financial Studies,* 21, 1455–508.

Xiao, Z (2009). "Quantile cointegrating regression," *Journal of Econometrics*, 150, 248-260.