Testing the Predictive Ability of Possibly Persistent Variables under Asymmetric Loss*

Matei Demetrescu† University of Kiel

Christoph Roling University of Bonn

Preliminary version: March 29, 2014

Abstract

We study predictability of economic time series under a general loss function. While for stationary predictors this task does not pose difficulties, non-standard limiting distributions of standard inference methods arise once the regressors are endogenous (i.e. there is contemporaneous correlation between shocks of the regressor and the dependent variable) and persistent (i.e. the predictor is reverting slowly to its long-run mean, if at all). Endogeneity turns out to be loss-function specific; thus, no endogeneity under MSE does not imply, and is not implied by, endogeneity under, say, an asymmetric quadratic loss function. Existent solutions for the endogeneity problem in predictive regressions with predictors of unknown persistence are valid for OLS-based inference only, and thus apply exclusively to MSE-optimal forecasting. We propose an overidentified instrumental variable-based test, using a highly trending, yet exogenous instrument and a possibly endogenous, yet less persistent one. The test statistic is analogous to the Anderson-Rubin statistic while taking the relevant loss into account; moreover, it follows a chi-squared distribution asymptotically irrespective of the degree of persistence of the predictor. The proposed methodology is applied with the forward premium puzzle by providing evidence for deviations from MSE loss and by conducting robust inference of the rational expectations hypothesis.

Key words: General cost-of-error function, Unknown persistence, Endogeneity, Robustness

JEL classification: C12 (Hypothesis Testing), C22 (Time-Series Models)

*The authors would like to thank Jörg Breitung und Sinem Hacioglu for helpful comments.

†**Corresponding author:** Institute for Statistics and Econometrics, Christian-Albrechts-University of Kiel, Olshausenstr. 40-60, D-24118 Kiel, Germany, email: mdeme@stat-econ.uni-kiel.de.

1 Introduction

Inference in predictive regressions is an ongoing topic in economics and finance. For instance, forward premium regressions test whether current forward rates are unbiased predictors of future spot exchange rates, while stock return regressions examine if economic fundamentals predict future stock returns.[1] With the exception of Maynard et al. (2011) and Lee (2012), who consider a quantile regression approach, the vast majority of this research is confined to inference using ordinary least squares (OLS) estimation. Therefore, existing analyses adopt the mean squared error (MSE) criterion to construct forecasts and, consequently, test the rational expectations or efficient market hypotheses in an MSE framework.

There is a significant body of evidence, however, that forecasters do not rely exclusively on MSE as criterion for forecast optimality. In a macroeconomics, Artis and Marcellino (2001) find systematic over- and underpredictions in IMF and OECD forecasts of the deficit in G7 countries. Elliott et al. (2005) discuss a method to estimate the degree of asymmetry of a loss function; using this method, Christodoulakis and Mamatzakis (2008, 2009) analyze series of g.d.p. growth forecasts of EU institutions and countries to reveal asymmetric preferences of forecasters. In addition, Capistrán (2008) even finds evidence of time-varying asymmetric preferences. More recently, Pierdzioch et al. (2012b) find evidence of asymmetry in the loss function of the Bank of Canada, and Komunjer and Owyang (2012) extend the work of Elliott et al. (2005) to a multivariate setting. In finance, Clatworthy et al. (2012) and Aiolfi et al. (2010) argue that financial analysts bear different costs for over - or underpredicting firms' earnings and are hence likely to have an asymmetric loss function.

Using a certain loss function to obtain optimal forecasts leads to estimation under the relevant loss function (see Granger (1969) and Weiss (1996)); that is, to obtain an estimate of the forecast optimal under a given loss function, one should estimate relevant parameters by minimizing the observed loss. The reason to do so is illustrated by the difference between OLS and least absolute deviations (LAD) estimation of a predictive regression (cf. Maynard et al., 2011). Let a regression disturbance term have zero conditional mean given the regressor, yet be conditionally heteroskedastic. Under MSE, the optimal prediction is zero and the predictor useless; an OLS-based test of predictability has power equal to size. Under LAD, however, the optimal prediction is the conditional median, which depends on the predictor via the conditional variance whenever the distribution of the shocks is not symmetric; see Christoffersen and Diebold (1997), for example. LAD estimation and testing will consequently detect predictability. Using the relevant loss function for estimation and subsequent predictability testing is therefore essential when evaluating the power to predict with respect to a given loss function.

[1] The empirical research in either of these areas is enormous. See Engel (1996) for economic background and early reviews of the empirical evidence of forward premium regressions, and Welch and Goyal (2008) for stock return regressions.

Two features characterize statistical inference in predictive regressions. First, it is often the case that the shocks occurring to the predictor and the dependent variable are contemporaneously correlated (the predictor is then called endogenous in the predictive regressions literature). Second, many predictors display (very) slow mean reversion, if at all (the predictor is then said to be persistent). It is this combination of endogeneity and persistence that invalidates standard OLS-based inference in predictive regressions: in case of nearly integrated regressors, Elliott and Stock (1994) show the distribution of the usual OLS t statistic to depend on both the degree of endogeneity and the persistence of the regressor.[2] If the regressor is stationary, however, the limiting null distribution is standard normal as expected. This discontinuity poses problems when the degree of persistence of the regressor is unknown: Cavanagh, Elliott, and Stock (1995) show pretesting to fail in the presence of nearly integrated regressors, for which the mean reversion parameter cannot be consistently estimated.

The complexity of the inferential problem gains an additional dimension under asymmetric loss. In case of stationary predictors, the asymptotics of estimation and testing under asymmetric loss (which can be cast as M estimation, cf. Huber, 1981) is standard, and poses no additional difficulties compared to quasi-maximum likelihood. A standard normal asymptotic null distribution emerges for the t statistic of the slope parameter in question, provided that weak regularity conditions on the loss function and the data generating process are fulfilled Amemyia (1985, Chapter 4). Empirical work, however, often investigates the predictive ability of economic variables that are persistent, such as forward premia (see e.g. Section 2.1). But how do estimators under the relevant loss behave in the presence of such stochastically trending variables? While it is expected that asymptotics similar to the OLS case arise even when estimating under an asymmetric loss (intuition confirmed by our asymptotic and small-sample simulation results) the relevant notion of endogeneity turns out to depend on the specific loss function. In a perhaps extreme, yet not unlikely scenario, there may be no endogeneity at all under OLS estimation, whereas the degree of endogeneity might be quite substantial under an asymmetric loss function.

Furthermore, the magnitude of the distortions depend on the type of persistence exhibited by the predictors. The workhorse model for persistent regressors has been the near unit root framework. Maynard and Phillips (2001) discuss cases where persistence can equally well be modelled in terms of a fractionally integrated process. As pointed out by Müller and Watson (2008), it is difficult to distinguish between the two persistent data generating processes in small samples; worse yet, they are not the only data generating processes exhibiting high persistence.[3]

Consequently, inferential tools developed for near-integrated regressors and OLS (see e.g. Camp-

[2] See also Stambaugh (1999). For a recent review of inference in predictive regressions, see Phillips and Lee (2013).

[3] Even a short-memory process with a break in the mean can mimic persistence; see among others Davidson and Sibbertsen (2005).

bell and Yogo, 2006, Jansson and Moreira, 2006, or Hjalmarsson, 2010) thus fail under asymmetric loss functions or other types of persistence. To bridge this gap in the literature we proceed as follows.

In section 2, an introductory empirical application is presented to point out statistical features that characterize the statistical testing problem as outlined above.

To provide correct inference, we draw in Section 3 on an instrumental variable (IV) approach as studied by Breitung and Demetrescu (2013) and propose a generalized M testing procedure that applies under asymmetric loss and that conveniently leads to a chi-square distribution under the null, irrespective of the degree and type of persistence of the predictor. To allow for these different possibilities in this respect, we consider a potential predictor to be persistent if the regressor, suitably normalized, converges weakly to a continuous-time process.

In Section 4 we reexamine the well-known forward premium puzzle. Using evidence for deviations from MSE loss for a collection of exchange rates, the rational expectations hypothesis is tested allowing for asymmetric loss functions. The testing procedure uses the robust IV approach and shows little evidence for failure of the rational expectations hypothesis.

The issue of testing forecast rationality under asymmetric loss functions has received attention in the literature. Batchelor and Peel (1998) for example test unbiasedness of forecasts under linex loss. Their analysis shows that the unbiasedness regression suffers from an omitted variable bias, as under the linex loss function, the optimal predictor not only depends on the conditional mean but also on the conditional variance of the dependent variable. They therefore suggest including a suitable estimate of the conditional variance as an additional regressor and to test forecast unbiasedness in this enriched regression. Patton and Timmermann (2007) study several data generating processes that govern the dynamics of the conditional mean and the conditional variance and test forecast rationality for a class of loss functions without imposing a particular parametric form. Aretz et al. (2011) complement the approaches suggested by Elliott et al. (2005) and Patton and Timmermann (2007) with a block-bootstrap to test forecast rationality using survey data. In contrast to existing contributions in the literature, this paper focuses on testing forecast rationality under an asymmetric loss function when the predictors are explicitly allowed to be persistent. This framework hence addresses several features relevant for practitioners, including asymmetric preferences about forecast errors and uncertainty about the degree of persistence of the potential predictors.

Before proceeding to the main part of this paper, let us introduce some notation. Let $\mathbf{1}(\cdot)$ denote the indicator function, $\mathbf{1}(A) = 1$ if proposition A is true and $\mathbf{1}(A) = 0$ otherwise. The lag operator is denoted by L, $L\{y_t\} = \{y_{t-1}\}$. The L_p norm of a random variable y_t is denoted as $||y_t||_p = (\mathbb{E}|y_t|^p)^{1/p}$. Weak convergence on a space of càdlàg functions endowed with a suitable norm is denoted by "$\Rightarrow$." Finally, "$\xrightarrow{p}$" stands for convergence in probability and "$\xrightarrow{d}$" stands for convergence in distribution. All proofs of the theorems are relegated to the appendix.

2 Asymmetric loss and forecast rationality in the foreign exchange market

2.1 Preliminaries

To illustrate some of the theoretical issues investigated in section 3 and to prepare the empirical results in section 4, it is instructive to study forecast rationality in the foreign exchange market. If $\mathbb{E}_t[S_{t+k}]$ denotes the spot price of a given currency that is expected to prevail at time $t+k$, it is natural to postulate $\mathbb{E}_t[S_{t+k}] = F_t^{t+k}$, where F_t^{t+k} denotes the k-period ahead forward exchange rate available at time t, and $\mathbb{E}_t[\cdot]$ is the conditional expectation with respect to the information available up to time t. Under the null hypothesis of rational expectations, then, in a regression of future spot rates on current forward rates, the coefficient attached to the forward rate is equal to one, see Geweke and Feige (1979), for example. The forward rate is then said to be an unbiased predictor of future spot rates. A more widely used empirical model to examine this issue does not consider the formulation in levels, but rather the changes in the spot rates as in

$$s_{t+k} - s_t = \gamma_0 + \gamma_1 \left(f_t^{t+k} - s_t\right) + u_{t+k}, \tag{1}$$

where s_t is the logarithm of a given spot exchange rate at time t, f_t^{t+k} is the logarithm of the forward exchange rate for time $t+k$ formed at time t and u_{t+k} is an idiosyncratic error; see, among others, Fama (1984). If agents are risk neutral, the rational expectations hypothesis corresponds to testing $\gamma_0 = 0$ and $\gamma_1 = 1$. A deviation from this pure form of the rational expectations hypothesis allows for an intercept different from zero and focuses on testing $\gamma_1 = 1$, and we follow this approach (see for example Liu and Maynard (2005)).

It is convenient in our case to consider the transformed regression

$$s_{t+k} - f_t^{t+k} = \beta_0 + \beta_1 \left(f_t^{t+k} - s_t\right) + u_{t+k}, \tag{2}$$

with $\beta_0 = \gamma_0$ and $\beta_1 = \gamma_1 - 1$. Accordingly, the hypothesis of interest is $\beta_1 = 0$. This regression can also be viewed as a test of the efficient markets hypothesis: if exchange rate markets are efficient, in the sense that market participants fully exploit all currently available information when forming expectations of future prices, then the forecast error $s_{t+k} - f_t^{t+k}$ is uncorrelated with any variable available at time t. Hence the coefficient β_1 is equal to zero; see Hansen and Hodrick (1980), for example.

A typical finding in the literature is that the estimated slopes in regression (1) differ significantly from one and often have a negative sign, which is therefore evidence against the rational expectations hypothesis; see Lewis (1995) for further details.

It should be noted that it is implicitly assumed that agents face a quadratic loss function. A different loss function implies that a test of the rational expectations hypothesis is conducted under the relevant loss, that is, the parameters in (1) or (2) are estimated taking into account the respective loss function, and hypothesis tests in these models are carried out using these estimates.

Recently, Christodoulakis and Mamatzakis (2013) find evidence for the so called Quad-Quad loss function in monthly exchange rate series of G7 countries. This loss function is given by

$$\mathcal{L}\left(s_{t+k} - f_t^{t+k}\right) = \left((1-2\alpha)\,\mathbf{1}\left(s_{t+k} - f_t^{t+k} < 0\right) + \alpha\right)|s_{t+k} - f_t^{t+k}|^2, \tag{3}$$

which boils down to MSE loss for $\alpha = 0.5$. It imposes a higher penalty for overprediction of the exchange rate, however, if $\alpha < 0.5$, while underprediction is more costly if $\alpha > 0.5$.

To investigate possible asymmetries, we consider weekly data for a collection of exchange rates, expressed as the price of foreign currency for one US dollar. These exchange rates are the end-of-week spot and one month forward rates taken from the Barclays Bank index and are obtained from Datastream. We study a four week horizon such that $k = 4$ in the above regressions. The sample period is 01/03/1992 - 05/24/2013. Given the evidence in Christodoulakis and Mamatzakis (2013) who present evidence for asymmetric loss in the post 2002 period, we also consider the subsample beginning on 01/08/2002.

Table 1 presents summary statistics of the spot and forward rates and the relevant regression variables used in (2). The first order autocorrelations of the predictor $f_t^{t+4} - s_t$ is large for many exchange series, in particular in the 2002 subsample, indicating that the forward premium is quite persistent in these cases.

2.2 Estimating loss function parameters

We follow Elliott, Komunjer, and Timmermann (2005) to estimate the parameter α in (3). The loss function parameter is estimated as

$$\widehat{\alpha} = \frac{\left[\frac{1}{T-5}\sum_{t=2}^{T-4} w_{t+3}\left|\, s_{t+4} - f_t^{t+4}\,\right|\right]' \widehat{S}^{-1}\left[\frac{1}{T-5}\sum_{t=2}^{T-4} w_{t+3}\mathbf{1}\left(s_{t+4} - f_t^{t+4} < 0\right)\left|\, s_{t+4} - f_t^{t+4}\,\right|\right]}{\left[\frac{1}{T-5}\sum_{t=2}^{T-4} w_{t+3}\left|\, s_{t+4} - f_t^{t+4}\,\right|\right]' \widehat{S}^{-1}\left[\frac{1}{T-5}\sum_{t=2}^{T-4} w_{t+3}\left|\, s_{t+4} - f_t^{t+4}\,\right|\right]},$$

where w_{t+3} is a vector of instruments, which are specified below, and T denotes the sample size. Here,

$$\widehat{S} = \frac{1}{T-5}\sum_{t=2}^{T-4} w_{t+3}w_{t+3}'\left(\mathbf{1}\left(s_{t+4} - f_t^{t+4} < 0\right) - \widehat{\alpha}\right)^2 \left|\, s_{t+4} - f_t^{t+4}\,\right|^2.$$

Table 1: Summary statistics for weekly exchange rate data (1992 - 2013)

		$s_{t+4} - f_t^{t+4}$	$f_t^{t+4} - s_t$	s_{t+4}	f_t^{t+4}
AUS					
	mean	-0.0025	0.0017	0.3003	0.3017
	std. dev.	0.0315	0.0015	0.1912	0.1905
	AC(1): 1992-2013	0.096	0.778	0.983	0.983
	AC(1): 2002-2013	-0.047	0.908	0.959	0.959
CAD					
	mean	-0.0005	0.0001	0.2315	0.2316
	std. dev.	0.0208	0.0011	0.1531	0.1529
	AC(1): 1992-2013	0.009	0.730	0.987	0.987
	AC(1): 2002-2013	-0.035	0.887	0.968	0.967
CHF					
	mean	-0.0002	-0.0011	0.2420	0.2409
	std. dev.	0.0302	0.0023	0.1815	0.1813
	AC(1): 1992-2013	0.057	0.621	0.981	0.981
	AC(1): 2002-2013	-0.097	0.945	0.954	0.954
EUR					
	mean	-0.0004	-0.0002	-0.1809	-0.1811
	std. dev.	0.0298	0.0014	0.1626	0.1626
	AC(1): 1999-2013	0.027	0.628	0.982	0.982
	AC(1): 2002-2013	0.013	0.945	0.933	0.933
GBP					
	mean	-0.0004	0.0011	-0.4942	-0.4932
	std. dev.	0.0266	0.0013	0.0929	0.0927
	AC(1): 1992-2013	0.067	0.725	0.956	0.956
	AC(1): 2002-2013	0.021	0.812	0.950	0.950
YEN					
	mean	0.0016	0.0023	4.6738	4.6715
	std. dev.	0.0302	0.0029	0.1445	0.1437
	AC(1): 1992-2013	0.002	0.462	0.976	0.975
	AC(1): 2002-2013	-0.065	0.973	0.973	0.973

Note: AC(1) denotes the first-order autocorrelation coefficient. The sample for EUR begins on 01/03/1999.

As the matrix $\widehat{S}$ depends on the estimated parameter, estimation is done iteratively, starting with $\widehat{S}$ as the identity matrix. Elliott, Komunjer, and Timmermann (2005) show that for a *given* pair of p and α, the optimal forecast ${f_t^*}^{t+k}$ under the above loss function satisfies the moment condition

$$\mathbb{E}\left[w_t\left(\mathbf{1}\left(s_{t+k} - {f_t^*}^{t+k} < 0\right) - \alpha\right) \mid s_{t+k} - {f_t^*}^{t+k} \mid^{p-1}\right] = 0,$$

and that the solution ${f_t^*}^{t+k}$ is uniquely characterized by this conditon. Conversely, then, for given forecasts, this condition can be used to solve for the asymmetry parameter, and the above estimator is the finite sample analogue of this solution. Furthermore, hypothesis test regarding $\widehat{\alpha}$ can be conducted using the limiting distribution of the estimated asymmetry parameter,

$$\begin{aligned}\sqrt{T}\left(\widehat{\alpha} - \alpha\right) &\xrightarrow{d} \mathcal{N}\left(0, V\right),\\ \left(\widehat{h}'\widehat{S}^{-1}\widehat{h}\right)^{-1} &\xrightarrow{p} V,\end{aligned}$$

with $\widehat{h} = 1/\left(T-5\right)\sum_{t=2}^{T-4} w_{t+3}|s_{t+4} - f_t^{t+4}|$. It is questionable whether the limiting distribution of $\widehat{\alpha}$ is indeed normal if some of the instruments, as for instance the forward rate, are persistent; see also Table 1. The limiting distribution of the estimated loss function parameter with persistent instruments is not available and may be subject to future research. We consider the limiting normal distribution as the best currently available approximation. Given that the differences between the estimates with possibly persistent instruments such as the forward rate and stationary instruments such as the forecast error point in a similar direction, the potential bias may considered as tolerable in this exercise.

Tables 2a and 2b present point estimates of the parameter α in (3) for the different series. The standard errors and probability values for testing the hypothesis $\alpha = 0$ against $\alpha \neq 0$ are also reported. The estimates are produced using four different sets of instruments, including the lagged spot rate, the lagged forward rate and the lagged forecast error. These instruments combine the choices made by Pierdzioch et al. (2012a), Christodoulakis and Mamatzakis (2013) and Elliott, Komunjer, and Timmermann (2005).

The estimates vary with the instruments employed. For the Australian Dollar, the results point towards a loss function parameter around 0.6 in the full sample and an even larger value in the 2002 subsample. Hence, somewhat surprisingly, the estimates suggests that underprediction of the exchange rate is more costly. A similar conclusion holds for the Canadian Dollar. For the Swiss Franc and the Yen, the loss function parameters are roughly estimated between 0.3 and 0.5 in the full sample. This range applies to the Yen in the 2002 - 2013 subsample as well, while the point estimates for the Swiss Franc are larger in this period. The evidence for the Euro and the British Pound is a bit more conflicting among the sets of instruments, with some of the estimates very close to the symmetric case in which $\alpha = 0.5$.

Table 2a: Loss function parameter estimates (Jan. 3 1992 - May 24 2013)

Instr.		AUS	CAD	CHF	EUR	GBP	YEN
I_1							
	$\widehat{\alpha}$	0.56	0.51	0.50	0.51	0.51	0.47
	s.e.	0.021	0.020	0.019	0.024	0.020	0.019
	prob. value	0.00	0.56	0.93	0.71	0.61	0.09
I_2							
	$\widehat{\alpha}$	0.65	0.59	0.43	0.63	0.63	0.33
	s.e.	0.018	0.020	0.019	0.022	0.019	0.017
	prob. value	0.00	0.00	0.00	0.00	0.00	0.00
I_3							
	$\widehat{\alpha}$	0.56	0.53	0.51	0.51	0.53	0.46
	s.e.	0.020	0.020	0.019	0.024	0.020	0.019
	prob. value	0.00	0.13	0.69	0.62	0.16	0.04
I_4							
	$\widehat{\alpha}$	0.65	0.57	0.40	0.63	0.63	0.33
	s.e.	0.018	0.019	0.018	0.022	0.019	0.017
	prob. value	0.00	0.00	0.00	0.00	0.00	0.00

Note: Four sets of instruments are used: a constant and the lagged spot rate (I_1), a constant and the lagged forecast error (I_2), a constant and the lagged forward rate (I_3), and a constant, the lagged spot rate, and the lagged forecast error (I_4). The p value is reported for the hypothesis test $\alpha = 0.5$ against $\alpha \neq 0.5$.

Table 2b: Loss function parameter estimates (Jan. 8 2002 - May 24 2013)

Instr.		AUS	CAD	CHF	EUR	GBP	YEN
I_1							
	$\widehat{\alpha}$	0.62	0.59	0.56	0.57	0.54	0.51
	s.e.	0.029	0.029	0.026	0.026	0.027	0.026
	prob. value	0.00	0.00	0.02	0.01	0.17	0.80
I_2							
	$\widehat{\alpha}$	0.73	0.68	0.70	0.74	0.68	0.42
	s.e.	0.023	0.025	0.024	0.022	0.024	0.026
	prob. value	0.00	0.00	0.00	0.00	0.00	0.00
I_3							
	$\widehat{\alpha}$	0.66	0.63	0.58	0.61	0.54	0.51
	s.e.	0.029	0.028	0.026	0.026	0.028	0.026
	prob. value	0.00	0.00	0.00	0.00	0.16	0.69
I_4							
	$\widehat{\alpha}$	0.75	0.69	0.68	0.74	0.68	0.41
	s.e.	0.023	0.024	0.024	0.02	0.024	0.025
	prob. value	0.00	0.00	0.00	0.00	0.00	0.00

Note: See the notes in Table 2a for additional information.

These results exemplify some of the features that may characterize rationality testing, in particular the presence of a possibly persistent predictor and the tendency of forecasters to care differently about positive and negative forecasting errors. In the next section, we study these properties in a theoretical framework to obtain a valid test of the rational expectations hypothesis in this case.

3 Estimation and inference in the predictive regression model under asymmetric loss

This section presents the main theoretical results. We first review the predictive regression model. In particular, a suitable notion of persistence is given and is distinguished from stationarity of possible predictors. Then, a particular class of asymmetric loss functions is introduced into this model; the class nests the MSE loss function that is typically adopted in predictive regression models. Third, estimation and inference in this framework is examined.

3.1 Model and statistical loss function

Consider the prototypical predictive regression model

$$y_t = \beta_0 + \beta_1 x_{t-1} + u_t, \tag{4}$$

where the null of interest is $\beta_1 = 0$. We discuss only the simple regression case in detail; as shall be seen, the extension to multiple regressors (of potentially different persistence) is straightforward and comes at virtually no additional cost.

The regressor x_{t-1} exhibits serial dependence, and is either highly persistent or stationary. To allow for a more precise definition of high persistence versus stationarity, we cast the data generating process in a time-varying linear process framework,

$$x_t = v_t + \sum_{j=1}^{t} \psi_{j,T} v_{t-j}. \tag{5}$$

The shocks u_t and v_t are taken to satisfy standard regularity conditions, see Assumption 2 below; in particular they are allowed to be contemporaneously dependent at time t to capture endogeneity in the predictive regression model (4). Deterministic components for the regressor can be introduced additively.

Depending on the values of the coefficients $\psi_{j,T}$, different behavior arises for the regressor in the limit. For instance, $\psi_{j,T} = (1 - c/T)^j$ leads to a nearly integrated regressor, while $\psi_{j,T} = \rho^j$

with $|\rho| < 1$ fixed leads to an asymptotically stationary regressor.[4] See Example 1 below. At the same time, short-run dynamics is allowed for; e.g. in the near-integrated case by letting $\psi_{j,T}$ be the convolution of a near-integrated AR(1) filter and a stationary AR component.

The framework allows for other data generating processes than fractional or near integration. We shall denote x_t as being highly persistent if the following definition is met.

Definition 1 *A process x_t is highly persistent in our framework if the coefficients $\psi_{j,T}$ in (5) are of such nature that*

(i) Δx_t *is uniformly L_2-bounded such that* $\sup_t ||\Delta x_t||_2 < C < \infty$.

(ii) there exists a sequence $n_T \to \infty$ satisfying $n_T/T \to 0$, and a continuous-time Gaussian process $X(s)$, continuous in quadratic mean, such that

$$\frac{1}{n_T} x_{[sT]} \Rightarrow \sigma_v X(s), \tag{6}$$

jointly with the convergence of the partial sums of u_t and v_t (regularity conditions on u_t, v_t provided).

(iii) $\limsup_{T\to\infty} \frac{1}{n_T^2} \sum_{j=1}^{T} \psi_{j,T}^2 < \infty$.

The analogy to the classical definition of an integrated process is quite strong: the differences of x_t are not trending, whereas the levels are nonstationary and nonergodic. At the same time, the weak limit of $x_{[sT]}$ is not restricted to be a Wiener process.

To deal with the case where the regressor is not highly persistent, we employ the following definition.

Definition 2 *A process x_t is weakly persistent in our framework if the coefficients $\psi_{j,T}$ in (5) are of such nature that*

$$\lim_{T\to\infty} \sum_{j=1}^{T} \psi_{j,T}^2 = C < \infty. \tag{7}$$

This condition ensures uniform L_2-boundedness of the regressor x_t in the limit and excludes trending behavior. Our derivations will rely on the above representations, so our findings apply whenever (6) or (7) holds.

[4] The term asymptotically stationary is used if the difference to a stationary process vanishes as $t \to \infty$; e.g. for fixed $|\rho| < 1$, $\sum_{j=1}^{\infty} \rho^j v_{t-j}$ is stationary and the difference to x_t, $\sum_{j=t+1}^{\infty} \rho^j v_{t-j}$, converges to zero in probability.

Example 1 *Let v_t be an iid sequence.*

1. *If x_t is generated according to $\psi_{j,T} = (1 - c/T)^j$ and $x_0 = o_p\left(\sqrt{T}\right)$, then*

$$\frac{1}{\sqrt{T}} x_{[sT]} \Rightarrow \sigma_v J_c(s),$$

with $J_c(s) = V(s) - c\int_0^s e^{-c(s-r)} V(r)\,\mathrm{d}r$ a standard Ornstein-Uhlenbeck (OU) process initialized at 0.

2. *If $\Delta_+^d x_t = v_t$, where $d \in (0.5, 1.5)$ and $\Delta_+^d = \mathbf{1}(t > 0)\Delta^d$ is the truncated version of the fractional difference operator given by the usual series expansion $(1 - L)^d = \Delta^d = \sum_{j\geq 0} \delta_j L^j$, then, with $B_d(s)$ a type-II fractional Brownian motion,*

$$\frac{1}{T^{d-0.5}} x_{[sT]} \Rightarrow \sigma_v B_d(s).$$

3. *If $\Delta_+^d x_t = v_t$ and $d < 0.5$, then x_t is asymptotically stationary.*

Let us now turn our attention to the loss function. According to Granger (1969), loss functions are quasi-convex functions minimized uniquely at zero. We shall adopt the more specific proposal of Elliott et al. (2005), and require that

Assumption 1 *The loss function $\mathcal{L}(u) \mapsto \mathbb{R}^+$ is given by*

$$\mathcal{L}(u) = ((1 - 2\alpha)\mathbf{1}(u < 0) + \alpha)|u|^p,$$

where $p \in \{2, 3, \ldots\}$ and $\alpha \in (0, 1)$.

Compared with Elliott et al. (2005), we do not consider the case $p = 1$ as it has already been discussed by Maynard et al. (2011).[5] The sign-based test proposed by Campbell and Dufour (1995) is in effect inference on the conditional median, and as such intimately related to LAD estimation.

The derivatives of the loss function will play an important role in the asymptotic analysis and in pinning down the notion of endogeneity. Assumption 1 makes $\mathcal{L}$ strictly convex and smooth with first-order derivative given by

$$\mathcal{L}^{(1)}(u) = p(\alpha - \mathbf{1}(u < 0))|u|^{p-1},$$

[5]Maynard et al. (2011) discuss a Bonferroni-based solution to the endogeneity problem under persistence.

and second-order derivative

$$\mathcal{L}^{(2)}(u) = p(p-1)\left((1-2\alpha)\mathbf{1}(u<0)+\alpha\right)|u|^{p-2}.$$

Note that $\mathcal{L}$, $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$ are homogenous of orders p, $p-1$ and $p-2$. The the $p-1$st-order derivative satisfies a uniform Lipschitz condition, while the pth-order derivative is discontinuous and bounded.

Moving on to estimation and inference in this model, it is natural to base predictability tests on estimation of (4)

$$y_t = \widehat{\beta}_0 + \widehat{\beta}_1 x_{t-1} + \widehat{u}_t, \tag{8}$$

with "$\widehat{\cdot}$" standing for estimates under the relevant loss $\mathcal{L}$, i.e.

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right)' = \underset{\left(\beta_0^*,\beta_1^*\right)'}{\arg\min} \sum_{t=2}^{T} \mathcal{L}\left(y_t - \beta_0^* - \beta_1^* x_{t-1}\right). \tag{9}$$

Note that the intercept $\widehat{\beta}_0$ in (8) would not converge to β_0 under a general loss function; it rather captures the mean together with the so-called forecast bias under the relevant loss (Granger, 1969). Regularity conditions provided (e.g. Assumption 2 above), it actually has as probability limit the M-measure of location (Huber, 1981) of the shocks u_t, so we may redefine without loss of generality

$$\beta_0 = \underset{\beta_0^*}{\arg\min}\ \mathrm{E}\left(\mathcal{L}\left(u_t - \beta_0^*\right)\right). \tag{10}$$

It is merely a shift that does not affect inference on β_1, as was already noted by McDonald and Newey (1988) in the context of M-estimation of linear regression models with iid disturbances.

The natural choice for a test of the null $\beta_1 = 0$ is the *t* statistic of β_1,

$$t_{\beta_1} = \frac{\widehat{\beta}_1}{s.e.\left(\widehat{\beta}_1\right)}, \tag{11}$$

where the standard error of $\widehat{\beta}_1$ is given by the usual "sandwich" estimator,

$$s.e.\left(\widehat{\beta}_1\right) = \sqrt{\left[B_T^{-1} M_T B_T^{-1}\right]_{2,2}},$$

with

$$B_T = \begin{bmatrix} \sum_{t=2}^{T} \mathcal{L}^{(2)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right) & \sum_{t=2}^{T} x_{t-1}\mathcal{L}^{(2)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right) \\ \sum_{t=2}^{T} x_{t-1}\mathcal{L}^{(2)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right) & \sum_{t=2}^{T} x_{t-1}^2\mathcal{L}^{(2)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right) \end{bmatrix},$$

and

$$M_T = \begin{bmatrix} \sum_{t=2}^{T} \left(\mathcal{L}^{(1)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right)\right)^2 & \sum_{t=2}^{T} x_{t-1}\left(\mathcal{L}^{(1)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right)\right)^2 \\ \sum_{t=2}^{T} x_{t-1}\left(\mathcal{L}^{(1)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right)\right)^2 & \sum_{t=2}^{T} x_{t-1}^2\left(\mathcal{L}^{(1)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right)\right)^2 \end{bmatrix}.$$

Having established the statistical loss function as well as the distinction between persistence and (asymptotic) stationarity, the following assumption summarizes the framework and is made for future reference.

Assumption 2 *The series y_t and x_t, $t = 1, \ldots, T$, are generated as in* (4) *and* (5) *such that either* (6) *(high persistence) or* (7) *(low persistence) hold true, where the sequence $(u_t, v_t)'$ is an iid sequence with finite moments of order $2p$. If $p = 2$, the distribution of u_t has no atom at β_0.*

Should x_t be stationary, usual M estimators inference can be shown to apply, and $\widehat{\beta}_1$ is $\sqrt{T}$-consistent and asymptotically normal distributed. The *t* statistic t_{β_1} is itself standard normally distributed under the null of no predictability, $\beta_1 = 0$. See e.g. Amemyia (1985, Chapter 4).

As shown in detail in appendix A, however, the limiting behavior of t_{β_1} is nonstandard if x_t is highly persistent whenever there is endogeneity. This discussion is analogous to the symmetric case, see Cavanagh et al. (1995). However, endogeneity is now loss-function specific. To see this, define the generalized forecast error as

$$\widetilde{u}_t = \mathcal{L}^{(1)}\left(u_t - \beta_0\right), \tag{12}$$

such that

$$\begin{pmatrix} \widetilde{u}_t \\ v_t \end{pmatrix} \overset{iid}{\sim} \left(0, \begin{pmatrix} \sigma_{\widetilde{u}}^2 & \sigma_{\widetilde{u}}\sigma_v\widetilde{\omega} \\ \sigma_{\widetilde{u}}\sigma_v\widetilde{\omega} & \sigma_v^2 \end{pmatrix}\right).$$

Under Assumption 2,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]} \begin{pmatrix} \widetilde{u}_t \\ v_t \end{pmatrix} \Rightarrow \begin{pmatrix} \sigma_{\widetilde{u}} & 0 \\ 0 & \sigma_v \end{pmatrix} \begin{pmatrix} \widetilde{W}(s) \\ V(s) \end{pmatrix},$$

jointly with (6) whenever x_t is highly persistent, where $\left(\widetilde{W}(s), V(s)\right)'$ is a bivariate Brownian motion with covariance matrix $\begin{pmatrix} 1 & \widetilde{\omega} \\ \widetilde{\omega} & 1 \end{pmatrix}$. As shown in appendix A, for the t-statistic in (11)

we have under Assumptions 1 and 2 that

$$t_{\beta_1} \xrightarrow{d} \frac{\int_0^1 X(s)\,\mathrm{d}\widetilde{W}(s) - \widetilde{W}(1)\int_0^1 X(s)\,\mathrm{d}s}{\sqrt{\int_0^1 X^2(s)\,\mathrm{d}s - \left(\int_0^1 X(s)\,\mathrm{d}s\right)^2}}$$

as $T \to \infty$. Notice that the actual distribution depends on the limit of x_t (and is e.g. different when x_t is fractionally or near integrated). Moreover, it turns out that endogeneity is only governed by the correlation $\omega = \mathrm{corr}(u_t, v_t)$ when the loss function is the squared-error one and need not coincide with the degree of endogeneity under the asymmetric loss function, which is governed by the parameter $\widetilde{\omega}$, the correlation between the generalized forecast error $\widetilde{u}_t$ and the shocks to the predictor v_t. Hence, in theory, no endogeneity under MSE loss does not imply absence of endogeneity under an asymmetric loss function and vice versa. The latter point is illustrated in section 3.3

3.2 Inference under uncertainty about persistence

As pointed out by Elliott and Stock (1994), standard OLS inference is invalid if the regressor x_t is endogenous and highly persistent at the same time. This result holds analogously under the loss function studied here, see Theorem 4 in Appendix A. A simple way out is offered by the variable addition approach as proposed by Toda and Yamamoto (1995) and Dolado and Lütkepohl (1996). But as emphasized by Breitung and Demetrescu (2013) for the OLS case, this leads to a severe power loss, reducing the convergence rate of $\widehat{\beta}_1$ to $\sqrt{T}$, and a similar argument can be made here.[6] To avoid such a loss, we follow Breitung and Demetrescu (2013) and resort to overidentified IV estimation and testing with suitably chosen instrument variables. We adapt in the following their Anderson-Rubin (AR) type statistic to M estimation and testing under the relevant loss in (9).

Concretely, Breitung and Demetrescu (2013) consider two types of instruments. The first replaces the highly persistent regressor x_{t-1} by a less persistent one, that is still correlated strongly enough with the original variable to qualify as a valid instrument. At the same time, the reduced persistence allows for standard limiting distributions. The second type is given by strictly exogenous but highly persistent instruments. In the nearly integrated framework, several ways to construct such instruments are offered. The first type includes e.g. the mildly integrated process $(1 - \gamma_T L)_+^{-1} \Delta x_{t-1}$ for $\gamma_T = 1 - a/T^\delta$, with $a > 0$, and $0 < \delta < 1$ first discussed by Phillips and Magdalinos (2007). This type of instrument is also studied by Lee (2012) for the quantile regression procedure when the predictors are nearly integrated. We refer to such type-I instruments as $z_{t-1,T}^{(I)}$. The second type of instruments $z_{t-1,T}^{(II)}$ should be statistically independent of u_t

[6] This is a tedious, yet straightforward extension of the results of appendix A.

to guarantee exogeneity, and the class includes, for example, randomly generated random walks or functions of (scaled) time.

Model (5) allows predictors to be either weakly or highly persistent (including nearly or fractionally integrated processes, for example), so we look for a pair of instruments that is correlated strongly with the predictor and is able to mimic the persistence of the process, irrespective of whether it is highly or weakly persistent. Assumption 3 (i) below defines the first instrument as Δx_{t-1}, which is a convenient choice for our purpose since the first differences of x_t are not themselves highly persistent according to our definition of the highly persistent predictor. While it is true that Breitung and Demetrescu (2013) allow for a wider class of type-I instruments, they are able to do so by assuming that x_t is near-integrated. Relaxing the assumptions on the persistence type of the potential predictor comes at the price of restricting the kind of type-I instruments we rigorously deal with.

Assumption 3 *Let the instruments be given by* $\mathbf{z}_{t,T} = \left(z_{t,T}^{(I)}, z_{t,T}^{(II)}\right)'$. *Then*

(i) $z_{t-1,T}^{(I)} = z_{t-1}^{(I)} = \Delta x_{t-1}$.

(ii) $\mathbb{E}\left|z_{t-1}^{(I)}\right|^4 < C < \infty$, *and* $\mathbb{E}\left[\mathcal{L}^{(2)}\left(u_t - \beta_0\right)\middle| z_{t-1}^{(I)}, z_{t-2}^{(I)}, \ldots\right] = \eta^2 < \infty$.

(iii) $z_{[rT],T}^{(II)} \Rightarrow Z\left(r\right)$, *jointly with* $T^{-1/2}\sum_{t=1}^{[rT]}\left(\widetilde{u}_t\, v_t\right)'$.

(iv) plim $T^{-1}\sum_{t=2}^{T}\left(1, \mathbf{z}_{t-1,T}'\right)'\left(1, \mathbf{z}_{t-1,T}'\right) = \Sigma_z$ *as* $T \to \infty$, *for some finite and positive definite matrix* Σ_z.

Several options are available for the second type of instruments, and to fix ideas, let

$$z_{t-1,T}^{(II)} = \sin\left(\pi\left(t-1\right)/2T\right).$$

Of course, other choices are possible, but the sine function above is the leading term in a Loève-Karhunen expansion of X; see Phillips (1998), which gives an intuition about the identification mechanism when x_t is highly persistent. Following Breitung and Demetrescu (2013), the use of instruments of both types makes sure that identification occurs in the highly as well as the weakly persistent case.

With these choices, we test predictability using the Anderson-Rubin (AR) type statistic below. Intuitively, the test statistic checks whether generalized forecast errors under the null are correlated with the potential predictor, but does so by means of instruments. In this respect we are building on the work of Elliott et al. (2005). Concretely, the predictability test is conducted with

$$\mathcal{T} = \left(\sum_{t=2}^{T}\widetilde{\mathbf{z}}_{t-1,T}'\mathcal{L}^{(1)}\Big(y_t - \widehat{\beta}_0\Big)\right)\left(\sum_{t=2}^{T}\widetilde{\mathbf{z}}_{t-1,T}\widetilde{\mathbf{z}}_{t-1,T}'\left(\mathcal{L}^{(1)}\Big(y_t - \widehat{\beta}_0\Big)\right)^2\right)^{-1}\left(\sum_{t=2}^{T}\widetilde{\mathbf{z}}_{t-1,T}\mathcal{L}^{(1)}\Big(y_t - \widehat{\beta}_0\Big)\right), \tag{13}$$

where $\widehat{\beta}_0$ is the estimator of β_0 under the null hypothesis $\beta_1 = 0$,

$$\widehat{\beta}_0 = \arg\min_{\beta_0^*} \sum_{t=2}^{T} \mathcal{L}\left(y_t - \beta_0^*\right),$$

and $\widetilde{\mathbf{z}}_{t-1,T}$ is a 2×1 vector

$$\widetilde{\mathbf{z}}_{t-1,T} = \begin{bmatrix} z_{t-1,T}^{(I)} - \frac{1}{T}\sum_{t=2}^{T} z_{t-1,T}^{(I)} \\ z_{t-1,T}^{(II)} - \frac{1}{T}\sum_{t=2}^{T} z_{t-1,T}^{(II)} \end{bmatrix}.$$

One may obviously resort to more than just one instrument of each type without essentially changing the argument. Note however that additional instruments need not automatically improve the power of the test procedure, while the critical value of the AR statistic increases with the number of instruments. From an a priori perspective it seems preferable to stick to a small number of instruments – the more promising ones.

The properties of the statistic under the null are summarized in the following theorem.

Theorem 1 *Under Assumptions (1) - (3) and the null hypothesis, as* $T \to \infty$*,*

$$\mathcal{T} \xrightarrow{d} \chi^2(2),$$

irrespective of whether x_{t-1} *is weakly or highly persistent.*

Hence the AR statistic provides valid inference under asymmetric loss that is robust to the degree of persistence of the predictors. As Phillips and Lee (2013) emphasize, in predictive regressions it is of further interest to investigate the distribution of the test statistic if the null hypothesis does not hold to examine the ability of the test to detect predictability if it is indeed there. To this end, the local asymptotic power of the AR statistic is studied. Depending on the persistence of the predictor, we consider the sequence of alternatives

$$H_{1,T}: \ \beta_1 = \frac{b}{n_T\sqrt{T}} \tag{14}$$

for highly persistent regressors, and

$$H_{1,T}: \ \beta_1 = \frac{b}{\sqrt{T}} \tag{15}$$

for weakly persistent regressors. We obtain the following result.

Theorem 2 (i) *If the predictor* x_{t-1} *is persistent, then under Assumptions 1 - 3 and the sequence of local alternatives (14), as* $T \to \infty$*,*

$$\mathcal{T} \xrightarrow{d} \chi^2(2, \lambda_p),$$

with non-centrality parameter

$$\lambda_p = b^2 \frac{\left(\widetilde{\kappa}^{(2)}\right)^2 \sigma_v^2}{\sigma_{\widetilde{u}}^2 \left(\lim_{T\to\infty} \frac{1}{T} \sum_{t=2}^{T} \left(\widetilde{z}_{t-1}^{(II)}\right)^2\right)} \left(\int_0^1 \widetilde{Z}(s) X(s) ds\right)^2,$$

where $\widetilde{Z}(\cdot) = Z(\cdot) - \int_0^1 Z(r) dr$*.*

(ii) *If the predictor* x_{t-1} *is stationary, then under Assumptions 1 - 3 and the sequence of local alternatives (15), as* $T \to \infty$*,*

$$\mathcal{T} \xrightarrow{d} \chi^2(2, \lambda_s).$$

with non-centrality parameter

$$\lambda_s = b^2 \frac{\left(\widetilde{\kappa}^{(2)} \sigma_v^2\right)^2}{\sigma_{\widetilde{u}}^2 \left(\lim_{T\to\infty} \frac{1}{T} \sum_{t=2}^{T} \mathbb{E}\left[\left(\widetilde{z}_{t-1}^{(I)}\right)^2\right]\right)} \left(\sum_{j=0}^{\infty} \left(\psi_{j,T}^2 - \psi_{j,T}\psi_{j+1,T}\right)\right)^2.$$

Thus, the test is powerful in a $n_T^{-1}T^{-1/2}$ neighborhood around the null hypothesis under persistence, and in a $T^{-1/2}$ neighborhood under stationarity. With a persistent predictor, local power is determined by the type II instrument while the type I instrument is asymptotically negligible. The converse result holds if the predictor is weakly persistent.

It should be stressed that the $n_T^{-1}T^{-1/2}$ neighbourhood is where the naive M test would have power (considering the convergence rate of the M estimator $\widehat{\beta}_1$, see Theorem 3 in Appendix A), should one be able to fix its size problem in the general setup of Definition 1. This is illustrated in the following subsection.

Before moving on to some finite-sample examinations, note that, considering multiple regressors of uncertain persistence, a joint predictability test is immediately conducted along the same lines; moreover, the type-II instrument only has to be used once, in contrast to the type-I instruments which are regressor-specific by construction.

3.3 Endogeneity under asymmetric loss

To illustrate how endogeneity affects inference under asymmetric loss, we study a simple predictive regression model with a highly persistent regressor. Consider the the following regression system,

$$\begin{aligned} y_t &= \beta x_{t-1} + u_t, \\ x_t &= \rho_T x_{t-1} + v_t, \end{aligned} \tag{16}$$

for $t = 1, \ldots, T$, with focus on a nearly integrated predictor characterized by $\rho_T = 1 - c/T$ for some small nonnegative constant c and $x_0 = 0$. We are interested in testing the null hypothesis $\beta = 0$ if a quadratic, but asymmetric loss function applies,

$$\mathcal{L}(u_t) = \left((1 - 2\alpha)\,\mathbf{1}(u_t < 0) + \alpha\right)|u_t|^2. \tag{17}$$

This framework directly extends a widely used empirical model to an asymmetric treatment of prediction errors and contrasts with inference under the standard, symmetric quadratic loss function which leads to OLS estimation and inference in a possibly endogenous regression system.

As mentioned in Section 3.1 (see also Appendix A), endogeneity under asymmetric loss is determined by the correlation parameter $\widetilde{\omega} = \mathbb{E}\left[\widetilde{u}_t v_t\right]$ which need not coincide with correlation between the disturbance terms in the linear model (16). To discuss this distinction, suppose u_t is characterized by multiplicative heteroskedasticity,

$$u_t = \sigma_t \epsilon_t = \sigma \sqrt{f(v_t)}\epsilon_t. \tag{18}$$

where ϵ_t and v_t are iid standard normally distributed and are independent of each other, and $f(\cdot)$ is a function to be specified below. Here, shocks to the potential predictor affect the volatility of the variable of interest, which could arise in a stock return predictability context, if current shocks to the dividend yield or interest rates affect the variability of the return series, say. However, this model does not intend to translate a particular empirical characteristic of a given time series, but serves rather as a stylized framework to examine endogeneity under symmetric and asymmetric loss.

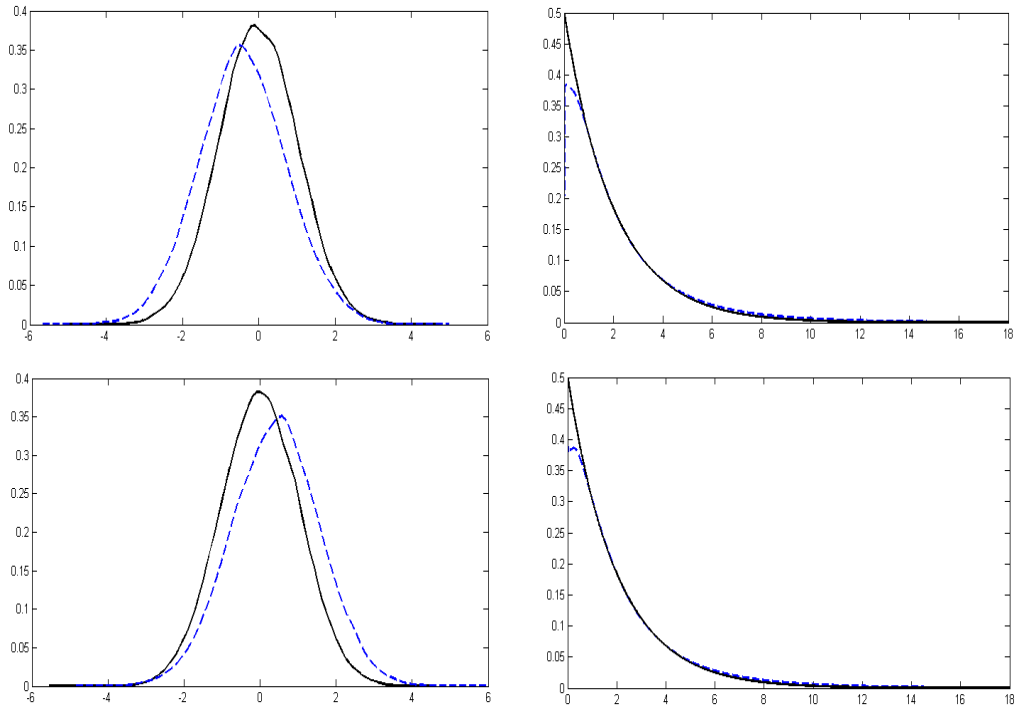
Clearly, (18) implies $\omega = \mathbb{E}\left[u_t v_t\right] = 0$, so even if the predictor of interest is persistent, standard inference applies in (16). In contrast, $\widetilde{\omega}$ may differ from zero, which, although analytically intractable, is exemplified for the following choice of $f(\cdot)$

$$f(v_t) = \left(|v_t| - \gamma v_t\right)^2, \tag{19}$$

which is an adaptation of the family of variance models suggested by Hentschel (1995). Here, the asymmetric nature of the shocks v_t implies that negative shocks receive a weight of $1+\gamma$, while a weight of $1-\gamma$ is assigned to positive shocks, where $|\gamma| \leq 1$. This asymmetric treatment allows positive and negative shocks of v_t to have a different impact on u_t as determined by the asymmetry parameter γ. The asymmetric response of many financial time series to positive and negative shocks to economic fundamentals has been documented (see, among others, Nelson (1991)), and the above specification incorporates this feature.

We generate 20,000 samples from this d.g.p. when $T = 100$, $c = 0$ and $\beta = 0$. We test predictive ability by computing the standard OLS t-statistic as well as the appropriate test statistic under asymmetric loss according to (11). In the asymmetric case, recall that $\mathcal{L}(u_t) = \mathcal{L}(y_t - \beta x_{t-1})$, such that, say, for $\beta \geq 0$, which arises naturally for predictors in return predictability studies, overprediction $(y_t < \beta x_{t-1})$ is more costly if $\alpha < 0.5$. For this exercise, we set $\alpha = 0.2$.

Figure 1: Densities of t-statistics for loss function parameter $\alpha = 0.2$



Note: The d.g.p. is given in (16) with $\rho_T = 1$, $\beta = 0$ and $T = 100$ and $f(v_t)$ as in (19). Top row: $\gamma = 0.95$, bottom row: $\gamma = -0.95$. Left column: estimated densities of the OLS t statistic (straigt line) and the t statistic under asymmetric loss (dotted line). Right column: densities of the $\chi^2(2)$ distribution (straight line) and the estimated density of the AR statistic (dotted line).

Figure 1 illustrates the consequences of an asymmetric loss function for inference in predictive regressions when $f(v_t)$ is given in (19) with $\gamma = 0.95$ (top row) and $\gamma = -0.95$ (bottom row). Regarding the left column of the figure, in absence of endogeneity the OLS t statistic is standard normally distributed, and the estimated density of the OLS t statistic (straight line) approaches

the density of the standard normal distribution (not shown). However, the induced error $\widetilde{u}_t = \mathcal{L}^{(1)}(u_t)$ and v_t may be correlated, and this correlation affects the asymptotic distribution of the t statistic under asymmetric loss (dotted line) as pointed out in Theorem 4 in Appendix A. For example, the combination of negative correlation between $\widetilde{u}_t$ and v_t with an integrated regressor shifts the density of the statistic under asymmetric loss to the right. A naive use of normal critical values for a one-sided test of the null hypothesis of $\beta = 0$ against $\beta > 0$ under the asymmetric quadratic loss function may lead a researcher to falsely reject the null hypothesis of no predictability. The AR statistic appears to provide inference with valid size in either case. The right column displays the densities of the AR statistic (dotted line) which approaches the density of the $\chi^2(2)$ distribution (straight line) as expected from Theorem 1.

To examine the size and power properties of the AR test in small samples, we conduct a second Monte Carlo experiment. We now consider two models for the predictors. First, the nearly integrated predictor is given by

$$x_t = \rho_T x_{t-1} + v_t \tag{20}$$

with $\rho_T = 1 - (c/T)$. Second, from the (truncated) representation of the fractionally integrated process $\Delta_+^d x_t = v_t$,

$$x_t = \sum_{j=0}^{t} \delta_j v_{t-j} \tag{21}$$

with $\delta_j = \Gamma(j+d) / (\Gamma(d)\Gamma(j+1))$. In either case, the dependent variable is generated as

$$y_t = \beta_0 + \beta_1 x_{t-1} + u_t \tag{22}$$

with $\beta_0 = 0$, $\beta_1 = b/T$, and where $(u_t, v_t)'$ are realizations of a bivariate normal distribution with mean vector equal to zero, unit variances and correlation $\omega = -0.9$. Therefore, the null hypothesis of no predictability corresponds $b = 0$. We consider the one-sided alternative $b > 0$. We assume the loss function in Assumption 1 applies with asymmetry parameter $\alpha = 0.25$.

For simplicity, the implied correlation parameter $\widetilde{\omega}$ is not reported, as ω and $\widetilde{\omega}$ are close in this exercise. The sample size is set as $T = 200$, nominal size is 0.05, and 10,000 replications are generated in each case.

Tables 3 and 4 present the results for the M test and the AR test statistic. In parallel to the symmetric MSE case, the M test is oversized, while the AR statistic provides valid inference for both the nearly and fractionally integrated models. The power loss of the robust test relative to the M test benchmark seems sizeable in this exercise, but it can be explained by the severe oversizedness of the M test.

Table 3: Size ($b = 0$) and size-adjusted power ($b > 0$) for the M test using t_{β_1} in (11)

nearly integrated predictor		$\rho_T = 1$	$\rho_T = 0.99$	$\rho = 0.97$	$\rho_T = 0.96$
	$b = 0$	0.36	0.25	0.16	0.14
	$b = 5$	0.58	0.44	0.26	0.22
	$b = 10$	0.92	0.95	0.81	0.70
	$b = 15$	0.96	1.00	0.99	0.98
	$b = 20$	0.96	1.00	1.00	1.00
fract. integrated predictor		$d = 1$	$d = 0.8$	$d = 0.6$	$d = 0.4$
	$b = 0$	0.20	0.20	0.14	0.10
	$b = 5$	0.79	0.27	0.13	0.10
	$b = 10$	0.99	0.75	0.31	0.19
	$b = 15$	1.00	0.96	0.57	0.32
	$b = 20$	1.00	1.00	0.81	0.48

Note: The data is generated according to (20) - (22) for a sample size of $T = 200$. The shocks $(u_t, v_t)'$ are drawn from a bivariate normal distribution with zero mean, unit variances and correlation equal to $\omega = -0.9$. The loss function in assumption 1 applies with asymmetry parameter $\alpha = 0.25$. For brevity, the implied correlation parameter $\widetilde{\omega}$ is not reported, as ω and $\widetilde{\omega}$ are close in this exercise. Nominal size is 0.05. The results are based on 10,000 replications.

Table 4: Size ($b = 0$) and size-adjusted power ($b > 0$) for the robust test using $\mathcal{T}$ in (13)

nearly integrated predictor		$\rho_T = 1$	$\rho_T = 0.99$	$\rho = 0.97$	$\rho_T = 0.96$
	$b = 0$	0.04	0.04	0.04	0.04
	$b = 5$	0.22	0.08	0.03	0.03
	$b = 10$	0.47	0.26	0.05	0.03
	$b = 15$	0.62	0.43	0.14	0.07
	$b = 20$	0.70	0.56	0.26	0.16
fract. integrated predictor		$d = 1$	$d = 0.8$	$d = 0.6$	$d = 0.4$
	$b = 0$	0.04	0.04	0.04	0.05
	$b = 5$	0.20	0.04	0.04	0.04
	$b = 10$	0.50	0.13	0.04	0.04
	$b = 15$	0.65	0.28	0.06	0.05
	$b = 20$	0.73	0.42	0.10	0.07

Note: see the notes to Table 3 for additional information.

4 Robust inference in forward premium regressions

Let us now return to testing the rational expectations hypothesis for the exchange rates in section 2. Combining the empirical evidence for asymmetric loss and persistent regressors allows us to test the rational expectations hypothesis with the inferential methods developed in the previous section. We restrict attention to the quadratic case $p = 2$ and allow $\alpha \neq 0.5$, which is a natural extension of the quadratic, symmetric loss functions considered in the literature.

We consider the regression

$$s_{t+k} - f_t^{t+k} = \beta_0 + \beta_1 \left(f_t^{t+k} - s_t \right) + u_{t+k}, \tag{23}$$

The hypothesis of interest is $\beta_1 = 0$.

Next, we estimate endogeneity in the regression system by

$$\widetilde{\omega} = \frac{\sigma_{\widetilde{u}v}}{\sigma_{\widetilde{u}}\sigma_v}, \tag{24}$$

with

$$\begin{aligned}
\sigma_v^2 &= \frac{1}{T}\sum_{t=2}^{T} \widehat{v}_t^{t+k}, \\
\sigma_{\widetilde{u}}^2 &= \frac{1}{T}\sum_{t=2}^{T} \left(\mathcal{L}^{(1)} \left(\widehat{u}_{t+k} - \widehat{\beta}_0 \right) \right)^2, \\
\sigma_{\widetilde{u}v} &= \frac{1}{T}\sum_{t=2}^{T} \mathcal{L}^{(1)} \left(\widehat{u}_{t+k} - \widehat{\beta}_0 \right) \widehat{v}_t^{t+k}.
\end{aligned}$$

Here, $\widehat{v}_t^{t+k}$ are the residuals from an estimated autoregressive model for the forward premium with lag selection by the BIC criterion, while $\widehat{u}_{t+k}$ are the residuals in (2). Table 5 reports the estimated correlation parameter for a range of loss function parameters α. For the Australian Dollar the correlation is moderately large, although for other series endogeneity may be less important. For comparison, inference is carried out with both the M (11) and the AR statistic.

Given these estimates of the loss functions, we investigate the rational expectations hypothesis under asymmetric, quadratic loss. The underlying regression is the transformed model (2) to test $\beta_1 = 0$. When this regression is actually run to carry out the test with standard inference, the dependent variable is constructed using non-overlapping time intervals to avoid serial correlation, and this approach is followed here as well when carrying out the test with the M-type t-statistic and the AR statistic. This results in 280 non-overlapping observations for the full sample and 149 observations for the subsample beginning in 2002. The AR statistic uses the first difference of the forward premium and the sine trend as instruments.

Table 5: Estimated correlation parameters $\widetilde{\omega}$

	Loss function parameter α				
	0.30	0.40	0.50	0.60	0.70
Jan. 3 1992 - May 24 2013					
AUS	-0.189	-0.199	-0.203	-0.203	-0.199
CAD	-0.032	0.027	-0.024	-0.020	-0.016
CHF	-0.062	-0.053	-0.049	-0.048	-0.050
EUR	-0.070	-0.070	-0.067	-0.062	-0.056
GBP	-0.150	-0.153	-0.155	-0.155	-0.152
YEN	-0.058	-0.066	-0.074	-0.082	-0.091
Jan. 8 2002 - May 24 2013					
AUS	-0.034	-0.034	-0.035	-0.36	-0.031
CAD	-0.144	-0.126	-0.106	-0.082	-0.056
CHF	-0.161	-0.145	-0.126	-0.104	-0.076
EUR	-0.186	-0.188	-0.185	-0.174	-0.156
GBP	-0.035	-0.034	-0.037	0.041	-0.048
YEN	-0.272	-0.251	-0.232	-0.212	-0.192

Note: The correlation parameter is estimated according to (24). The sample for EUR begins on 01/03/1999.

To accommodate for the variation in the point estimates from the different sets of instruments, the test is carried out for a range of values of the parameter α, where $\alpha = 0.5$ serves as a reference point in which inference is conducted that is robust to the degree of persistence of the regressor, and assumes a symmetric, quadratic loss function. For $\alpha \neq 0.5$, robust inference is made allowing for an asymmetric loss function. Tables 6 and 7 show the p values of test $\beta_1 = 0$ for the t statistic based on the asymptotic standard normal distribution (valid in absence of endogeneity) and the AR statistic based on the asymptotic $\chi^2(2)$ distribution. The results are reported for the Australian Dollar, the Canadian Dollar, the Swiss Franc and the Yen, as the symmetric loss function seems be a reasonable approximation for the Euro and the British pound from our earlier results.

Starting with the M test and taking the results in the symmetric case $\alpha = 0.5$ as a starting point, the rational expectations hypothesis is not rejected for the majority of the series. For the Australian and the Canadian Dollar, some further comments can be made in the full sample. First, Table 5 provides evidence for correlation in the regression system for the Australian Dollar, which leads to biased inference using the M test. For the Canadian Dollar, the estimated correlation is smaller and the M test may thus be taken to yield valid inference. The null is rejected in the symmetric case. Given the evidence for asymmetric loss with an estimated loss function parameter of about 0.55 or larger, the null is barely rejected or not rejected in these cases.

Table 6: p values for the test of $\beta_1 = 0$ using the M-type t-statistic

	Loss function parameter α								
	0.30	0.35	0.40	0.45	**0.50**	0.55	0.60	0.65	0.70
Jan. 3 1992 - May 24 2013									
AUS	0.01	0.01	0.02	0.03	**0.04**	0.07	0.10	0.16	0.24
CAD	0.06	0.05	0.04	0.04	**0.04**	0.05	0.06	0.09	0.12
CHF	0.11	0.14	0.15	0.19	**0.23**	0.28	0.33	0.39	0.48
YEN	0.99	0.85	0.73	0.62	**0.53**	0.45	0.39	0.34	0.29
Jan. 8 2002 - May 24 2013									
AUS	0.85	0.96	0.93	0.84	**0.76**	0.69	0.61	0.54	0.46
CAD	0.88	0.91	0.94	0.96	**0.98**	0.98	0.97	0.93	0.89
CHF	0.32	0.32	0.32	0.33	**0.36**	0.39	0.45	0.52	0.60
YEN	0.79	0.78	0.79	0.81	**0.85**	0.90	0.96	0.97	0.86

Note: The t statistic is given in (11).

Table 7: p values for the test of $\beta_1 = 0$ using the AR statistic

	Loss function parameter α								
	0.30	0.35	0.40	0.45	**0.50**	0.55	0.60	0.65	0.70
Jan. 3 1992 - May 24 2013									
AUS	0.01	0.01	0.03	0.05	**0.10**	0.17	0.26	0.38	0.52
CAD	0.00	0.00	0.01	0.01	**0.02**	0.04	0.07	0.12	0.18
CHF	0.98	0.96	0.91	0.84	**0.78**	0.72	0.68	0.64	0.61
YEN	0.71	0.78	0.84	0.88	**0.91**	0.93	0.93	0.94	0.93
Jan. 8 2002 - May 24 2013									
AUS	0.64	0.63	0.61	0.57	**0.53**	0.50	0.45	0.41	0.36
CAD	0.46	0.40	0.36	0.34	**0.32**	0.31	0.30	0.28	0.27
CHF	0.75	0.76	0.77	0.77	**0.76**	0.76	0.74	0.71	0.67
YEN	0.88	0.86	0.84	0.81	**0.76**	0.71	0.64	0.56	0.47

Note: The AR statistic is given in (13).

Next, considering the results for the AR statistic, for the Canadian Dollar, the null hypothesis $\beta_1 = 0$ is rejected at the 5 % level when $\alpha = 0.5$. Some of the estimation results of Table 2a suggests that the loss function parameter may range between 0.55 and 0.60. For these specifications, the null hypothesis is barely rejected or not rejected by the test using the AR statistic, such that evidence against the rational expectations hypothesis is weaker for this range.

Hence in this example, even after taking the uncertainty about the degree of persistence into account, different conclusions are reached for the symmetric case and plausible asymmetric

specifications. A similar observation can be made for the Australian Dollar. The p values for $\alpha = 0.45$ and $\alpha = 0.5$ suggest that the null hypothesis may be rejected or barely not rejected at the 10% level, while the results for the range of α parameters between 0.55 and 0.65 agree with the rational expectation hypothesis, and this range arises from the estimation results for the Australian Dollar in Table 2a.

5 Concluding remarks

We extend the linear predictive regression model to incorporate asymmetric loss functions, with the standard mean squared error (MSE) loss as a special case. The main interest is to test whether current observations of included regressors have predictive ability about the outcome variable in the next period. As in the standard case, the distribution of the t statistic based on M estimation under the relevant loss depends on the persistence of the predictor and the correlation between shocks to the predictor and the dependent variable. In contrast to the standard case, however, endogeneity depends on the adopted loss function and need not coincide with endogeneity under MSE loss. Hence in some cases the OLS t statistic may be standard normally distributed, while the t statistic under asymmetric loss has a non-standard distribution and vice versa. In addition, as the degree of persistence of the predictor is difficult to determine precisely, a test statistic is introduced in this setup that allows to conduct inference using the χ^2 distribution whether the predictor is weakly or highly persistent in a quite general sense. The predictive regression model under asymmetric loss is employed to investigate the forward premium puzzle for a collection of currencies. In these time series, a tendency for asymmetric treatment of overpredictions and underpredictions of future spot rates by forward rates is provided. Given these estimates of the loss function parameters, predictability is tested with the test statistic that is robust to the degree of persistence of the forward premium, and there appears little evidence for failure of the rational expectations hypothesis.

Appendix

A Asymptotics in the highly persistent case

In the nonstationary case, the behavior of the estimators under the relevant loss parallels that of the OLS estimators under near or fractional integration, and $\widehat{\beta}_1$ is consistent with a convergence rate depending on the persistence of the regressor, as indicated by the following theorem giving the asymptotic distributions of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. It becomes clear from the exposition, however, that endogeneity is only governed by the correlation $\omega = \operatorname{corr}(u_t, v_t)$ when the loss function is the squared-error one, and the general condition depends on the given loss function. So let

$$\widetilde{u}_t = \mathcal{L}^{(1)}(u_t - \beta_0), \tag{25}$$

such that

$$\begin{pmatrix} \widetilde{u}_t \\ v_t \end{pmatrix} \overset{iid}{\sim} \left(0, \widetilde{\Sigma}\right),$$
$$\widetilde{\Sigma} = \begin{pmatrix} \sigma_{\widetilde{u}} & 0 \\ 0 & \sigma_v \end{pmatrix} \begin{pmatrix} 1 & \widetilde{\omega} \\ \widetilde{\omega} & 1 \end{pmatrix} \begin{pmatrix} \sigma_{\widetilde{u}} & 0 \\ 0 & \sigma_v \end{pmatrix}.$$

(The fact that $\widetilde{u}_t$ has zero expectation comes form the fact that β_0 has been redefined as the M-measure of location of u_t under $\mathcal{L}$.) Under Assumption 2,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[sT]} \begin{pmatrix} \widetilde{u}_t \\ v_t \end{pmatrix} \Rightarrow \begin{pmatrix} \sigma_{\widetilde{u}} & 0 \\ 0 & \sigma_v \end{pmatrix} \begin{pmatrix} \widetilde{W}(s) \\ V(s) \end{pmatrix},$$

jointly with (6) whenever x_t is highly persistent, where $\left(\widetilde{W}(s), V(s)\right)'$ is a bivariate Brownian motion with covariance matrix $\begin{pmatrix} 1 & \widetilde{\omega} \\ \widetilde{\omega} & 1 \end{pmatrix}$.

Also, let $\widetilde{\kappa}^{(2)} = \mathrm{E}\left(\mathcal{L}^{(2)}(u_t - \beta_0)\right)$ and note that $\widetilde{\kappa}^{(2)} > 0$ due to the strict convexity of $\mathcal{L}$.

Theorem 3 *Under Assumptions 1 and 2, as $T \to \infty$,*

$$\sqrt{T}\left(\widehat{\beta}_0 - \beta_0\right) \xrightarrow{d} \frac{\sigma_{\widetilde{u}}}{\widetilde{\kappa}^{(2)}} \frac{\widetilde{W}(1) \int_0^1 X^2(s)\,\mathrm{d}s - \int_0^1 X(s)\,\mathrm{d}s \int_0^1 X(s)\,\mathrm{d}\widetilde{W}(s)}{\int_0^1 X^2(s)\,\mathrm{d}s - \left(\int_0^1 X(s)\,\mathrm{d}s\right)^2}$$
$$n_T\sqrt{T}\left(\widehat{\beta}_1 - \beta_1\right) \xrightarrow{d} \frac{\sigma_{\widetilde{u}}}{\widetilde{\kappa}^{(2)}\sigma_v} \frac{\int_0^1 X(s)\,\mathrm{d}\widetilde{W}(s) - \widetilde{W}(1) \int_0^1 X(s)\,\mathrm{d}s}{\int_0^1 X^2(s)\,\mathrm{d}s - \left(\int_0^1 X(s)\,\mathrm{d}s\right)^2}.$$

The persistence of the regressor increases, expectedly, the convergence rate of the estimator $\widehat{\beta}_1$. Also, $\sqrt{T}$-consistency of $\widehat{\beta}_0$ follows, although its distribution is nonstandard too.

In what concerns the main interest when testing the predictive power, the *t* statistic of β_1, we have the following result.

Theorem 4 *Under the assumptions of Theorem 3 we have for t_{β_1} from 11 as $T \to \infty$ that*

$$t_{\beta_1} \xrightarrow{d} \frac{\int_0^1 X(s)\,\mathrm{d}\widetilde{W}(s) - \widetilde{W}(1)\int_0^1 X(s)\,\mathrm{d}s}{\sqrt{\int_0^1 X^2(s)\,\mathrm{d}s - \left(\int_0^1 X(s)\,\mathrm{d}s\right)^2}}.$$

Remark 1 For $\mathcal{L}(u) = u^2$ and near integration, the usual distribution of the OLS-based *t* statistic established by Elliott and Stock (1994) is recovered. If $\widetilde{\omega} = 0$, the numerator is mixed Gaussian and the distribution of t_{β_1} is standard normal irrespective of the type of persistence x_t exhibits. Otherwise, the distribution may depend on nuisance parameters, e.g. when X is an OU process (where the nuisance parameter is the mean reversion parameter – which cannot be consistently estimated).

Remark 2 If $v_t \equiv u_t$ and $\psi_{j,T} = 1$, the distribution derived by Lucas (1995) for M estimation of unit root processes are obtained.

Remark 3 The quantile regression result of Lee (2012) for near integration can formally be derived from Theorem 4 using the method of Phillips (1991) and letting X be an OU process.

Remark 4 Extensions allowing for other types of deterministic components are straightforward. The distributions remain nonstandard as long as there is non-zero correlation between $\widetilde{u}_t$ and v_t and high persistence.

Since endogeneity is given in terms of correlation of v_t and $\widetilde{u}_t$ rather than in terms of correlation of v_t and u_t, we may encounter situations where endogeneity is not an issue if the loss function is of suitable nature. But this is not a guarantee, in fact chances are that endogeneity remains a problem as long as u_t and v_t are not independent.

B Proofs

Preliminary results

Lemma 1 *Let Assumptions 1 and 2 hold true. As $T \to \infty$, the following properties hold:*

1.

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}\begin{pmatrix}\widetilde{u}_t\\ v_t\end{pmatrix}\Rightarrow\begin{pmatrix}\sigma_{\widetilde{u}}\widetilde{W}(s)\\ \sigma_v V(s)\end{pmatrix},$$

where the standard Wiener processes V and $\widetilde{W}$ correlate with correlation $\widetilde{\omega}=\operatorname{corr}(\tilde{u}_t,v_t)$.

2. Furthermore, under persistence,

$$\frac{1}{n_T\sqrt{T}}\sum_{t=2}^{T}x_{t-1}\widetilde{u}_t\Rightarrow\sigma_v\sigma_{\widetilde{u}}\int_0^1 X(s)\,\mathrm{d}\widetilde{W}(s).$$

3. $\sup_t \mathrm{E}\left(x_{t-1}^{2p}\right)<\infty$.

4. Under persistence, for all $1\leq k\leq p$

$$\frac{1}{n_T^k T}\sum_{t=2}^{T}x_{t-1}^k\mathcal{L}^{(2)}(u_t-\beta_0)\Rightarrow\widetilde{\kappa}^{(2)}\sigma_v^k\int_0^1 X^k(s)\,\mathrm{d}s.$$

5. Similarly,

$$\frac{1}{n_T^k T}\sum_{t=2}^{T}x_{t-1}^k\widetilde{u}_t^2\Rightarrow\sigma_{\widetilde{u}}^2\sigma_v^k\int_0^1 X^k(s)\,\mathrm{d}s.$$

6. Finally, for any $\widetilde{\beta}_0=\beta_0+o_p(1)$, $\widetilde{\beta}_1=\beta_1+o_p\left(n_T^{-1}\right)$, *and* $k=0,1,2$, *we have under persistence that*

$$\frac{1}{n_T^k T}\sum_{t=2}^{T}x_{t-1}^k\mathcal{L}^{(2)}\left(y_t-\widetilde{\beta}_0-\widetilde{\beta}_1x_{t-1}\right)=\frac{1}{n_T^k T}\sum_{t=2}^{T}x_{t-1}^k\mathcal{L}^{(2)}(u_t-\beta_0)+o_p(1).$$

Proof of Lemma 1

1. obvious and omitted.
2. Follows from 1. given that $\widetilde{u}_t$ is independent of $v_{t-j}\ \forall j>0$; see Kurtz and Protter (1991).
3. We have that

$$\mathrm{E}\left(x_{t-1}^{2p}\right)=\sum_{j_1=1}^{t}\cdots\sum_{j_{2p}=1}^{t}\psi_{j_1,T}\cdots\psi_{j_{2p},T}\,\mathrm{E}\left(v_{t-j_1}\cdots v_{t-j_{2p}}\right);$$

given the zero-mean iid property of v_t, the indices $j_1,\ldots j_{2p}$ must be pairwise equal for the expectation on the right-hand side (r.h.s.) to be nonzero, so

$$\mathrm{E}\left(x_{t-1}^{2p}\right)=\sum_{j_1=1}^{t}\cdots\sum_{j_p=1}^{t}\psi_{j_1,T}^2\cdots\psi_{j_p,T}^2\,\mathrm{E}\left(v_{t-j_1}^2\cdots v_{t-j_p}^2\right).$$

Now, the expectation on the r.h.s. is uniformly bounded since v_t is iid with finite moments of order $2p$, so

$$\mathrm{E}\left(x_{t-1}^{2p}\right) \leq C\left(\sum_{j=1}^{t}\psi_{j,T}^{2}\right)^{p},$$

where the r.h.s. is uniformly bounded thanks to Definition 1.

4. Recall that $\widetilde{\kappa}^{(2)} = \mathrm{E}\left(\mathcal{L}^{(2)}\left(u_t - \beta_0\right)\right)$ and write

$$\frac{1}{n_T^k T}\sum_{t=1}^{T} x_{t-1}^{k}\mathcal{L}^{(2)}\left(u_t-\beta_0\right) = \widetilde{\kappa}^{(2)}\frac{1}{n_T^k T}\sum_{t=1}^{T} x_{t-1}^{k} - \frac{1}{n_T^k T}\sum_{t=1}^{T} x_{t-1}^{k}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right).$$

The result follows if the second summand on the r.h.s. vanishes as $T\to\infty$. But this is indeed the case. Note that $x_{t-1}^{k}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)$ is a martingale difference sequence given the iid property of $\left(u_t, v_t\right)'$ and thus of $\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}, v_t\right)'$, so

$$\mathrm{Var}\left(\frac{1}{n_T^k T}\sum_{t=1}^{T} x_{t-1}^{k}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)\right) = \frac{1}{n_T^{2k}T^2}\sum_{t=1}^{T}\mathrm{Var}\left(x_{t-1}^{k}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)\right).$$

The variances on the r.h.s. satisfy again due to the assumed iid property of the shocks

$$\mathrm{Var}\left(x_{t-1}^{k}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)\right) = \mathrm{E}\left(x_{t-1}^{2k}\right)\mathrm{E}\left(\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)^{2}\right),$$

where the first expectation is of order n_T^{2k} uniformly (given the finiteness of the moments of order $2p$ for v_t), and the second is uniformly bounded. The variance of the term $\frac{1}{n_T^k T}\sum_{t=1}^{T} x_{t-1}^{k}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)$ thus vanishes at rate T^{-1}.

5. Analogous to the proof of 4. and omitted.

6. Note first that, due to the weak convergence of x_t, we have

$$\sup_t\left(x_{t-1}\right) = O_p\left(n_T\right),$$

such that $\sup_t(\beta_1 - \widetilde{\beta}_1)x_t = o_p(1)$.

Then, for $p=3$, $\mathcal{L}^{(2)}$ is Lipschitz and the result follows immediately.

For $p>3$, use a Taylor expansion for $\mathcal{L}^{(2)}$ around $u_t-\beta_0$ to obtain

$$\begin{aligned}\mathcal{L}^{(2)}\left(y_t-\widetilde{\beta}_0-\widetilde{\beta}_1 x_{t-1}\right) = {} & \textstyle\sum_{j=2}^{p-2}\frac{1}{(j-2)!}\mathcal{L}^{(j)}\left(u_t-\beta_0\right)\left(\beta_0-\widetilde{\beta}_0+\left(\beta_1-\widetilde{\beta}_1\right)x_{t-1}\right)^{j-2} \\ & + \tfrac{1}{(p-3)!}\mathcal{L}^{(p-1)}\left(y_t-\widetilde{\beta}_0^{*}-\widetilde{\beta}_1^{*}x_{t-1}\right)\left(\beta_0-\widetilde{\beta}_0+\left(\beta_1-\widetilde{\beta}_1\right)x_{t-1}\right)^{p-3},\end{aligned}$$

for some $\widetilde{\beta}_0^{*}$ between β_0 and $\widetilde{\beta}_0$, and some $\widetilde{\beta}_1^{*}$ between β_1 and $\widetilde{\beta}_1$ (which implies $\widetilde{\beta}_0^{*}-\beta_0 = o_p\left(1\right)$ and $\widetilde{\beta}_1^{*}-\beta_1 = o_p\left(n_T^{-1}\right)$).
The leading term of the expansion ($j=2$) gives the desired r.h.s.; furthermore,

$$\sup_t\left(\beta_0-\widetilde{\beta}_0+\left(\beta_1-\widetilde{\beta}_1\right)x_{t-1}\right)^{j-2} = o_p\left(1\right).$$

Now,

$$\left|\frac{1}{n_T^k T}\sum_{t=2}^{T} x_{t-1}^k \mathcal{L}^{(j)}\left(u_t-\beta_0\right)\left(\beta_0-\widetilde{\beta}_0+\left(\beta_1-\widetilde{\beta}_1\right)x_{t-1}\right)^j\right|$$
$$\leq \frac{\sup_t\left(\beta_0-\widetilde{\beta}_0+\left(\beta_1-\widetilde{\beta}_1\right)x_{t-1}\right)^j}{n_T^k T}\sum_{t=2}^{T}\left|x_{t-1}^k \mathcal{L}^{(j)}\left(u_t-\beta_0\right)\right| = o_p\left(1\right).$$

For the last term of the expansion,

$$\frac{1}{n_T^{-k} T}\sum_{t=2}^{T} x_{t-1}^k \mathcal{L}^{(p-1)}\left(y_t-\widetilde{\beta}_0^*-\widetilde{\beta}_1^* x_{t-1}\right)\left(\beta_0-\widetilde{\beta}_0+\left(\beta_1-\widetilde{\beta}_1\right)x_{t-1}\right)^{p-1}.$$

Recall that $\mathcal{L}^{(p-1)}$ is Lipschitz, so

$$\left|\mathcal{L}^{(p-1)}\left(y_t-\widetilde{\beta}_0^*-\widetilde{\beta}_1^* x_{t-1}\right)-\mathcal{L}^{(p-1)}\left(u_t-\beta_0\right)\right| \leq C\left|\beta_0-\widetilde{\beta}_0^*+\left(\beta_1-\widetilde{\beta}_1^*\right)x_{t-1}\right|,$$

and the same reasoning as above applies, leading to the desired result for $p>2$.

For $p=2$, $\mathcal{L}^{(2)}$ is piecewise constant but discontinuous at 0 when $\alpha \neq 0.5$. Let $\xi_T = \widetilde{\beta}_0-\beta_0+\left(\widetilde{\beta}_1-\beta_1\right)x_{t-1}$ and note that $\xi_T \xrightarrow{p} 0$. We then have

$$\mathcal{L}^{(2)}\left(y_t-\widetilde{\beta}_0-\widetilde{\beta}_1 x_{t-1}\right) = \mathcal{L}^{(2)}\left(u_t-\beta_0-\xi_T\right)\left(\mathbf{1}\left(|\xi_T|\geq|u_t-\beta_0|\right)+\mathbf{1}\left(|\xi_T|<|u_t-\beta_0|\right)\right),$$

and note that $y_t-\widetilde{\beta}_0-\widetilde{\beta}_1 x_{t-1}$ can only switch sign when $|\xi_T|\geq|u_t-\beta_0|$. Thus, $\mathcal{L}^{(2)}\left(y_t-\widetilde{\beta}_0-\widetilde{\beta}_1 x_{t-1}\right) = \mathcal{L}^{(2)}\left(u_t-\beta_0\right)$ whenever $|\xi_T|<|u_t-\beta_0|$, and it suffices to show that

$$\frac{1}{n_T^k T}\sum_{t=2}^{T} x_{t-1}^k \mathcal{L}^{(2)}\left(y_t-\widetilde{\beta}_0-\widetilde{\beta}_1 x_{t-1}\right)\mathbf{1}\left(|\xi_T|\geq|u_t-\beta_0|\right)$$
$$\leq \sup_t \frac{x_{t-1}^k}{n_T^k}\sup_t \mathcal{L}^{(2)}\left(y_t-\widetilde{\beta}_0-\widetilde{\beta}_1 x_{t-1}\right)\frac{1}{T}\sum_{t=2}^{T}\mathbf{1}\left(|\xi_T|\geq|u_t-\beta_0|\right)\xrightarrow{p}0.$$

But $\sup_t \frac{x_{t-1}^k}{n_T^k} = O_p\left(1\right)$, and $\mathcal{L}^{(2)}$ is piecewise constant. Since $\mathrm{E}\left(\mathbf{1}\left(|\xi_T|\geq|u_t-\beta_0|\right)\right) = \Pr\left(|u_t-\beta_0|\leq|\xi_T|\right)$ vanishes when u_t does not have an atom at β_0, Markov's inequality implies that $\frac{1}{T}\sum_{t=2}^{T}\mathbf{1}\left(|\xi_T|\geq|u_t-\beta_0|\right)\xrightarrow{p}0$, as required for the result.

Lemma 2 *For $\widehat{\beta}_0$ and $\widehat{\beta}_1$ from* (9) *it holds under persistence as $T\to\infty$ that*

$$\left(\widehat{\beta}_0,\widehat{\beta}_1\right)' \xrightarrow{p} \left(\beta_0,\beta_1\right)',$$

such that

$$\widehat{\beta}_1-\beta_1 = o_p\left(n_T^{-1}\right).$$

Proof of Lemma 2

We begin by showing that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are consistent estimators, and establish the desired convergence rate in a second step.

A theorem of the type "if the target function converges uniformly in probability to deterministic function, minimized at the true values of the parameters, then argmin estimators are consistent" is used; see Chapter 4 of Amemiya (1985).

In order to establish the consistency of $\widehat{\beta}_1$, we distinguish two cases.

1. Let $\beta_1^* = \beta_1$. Then,

$$\frac{1}{T}\sum \mathcal{L}\left(y_t - \beta_0^* - \beta_1^* x_{t-1}\right) = \frac{1}{T}\sum \mathcal{L}\left(u_t - \beta_0^*\right) \xrightarrow{p} \mathrm{E}\left(\mathcal{L}\left(u_t - \beta_0^*\right)\right),$$

   pointwise in β_0^*, due to the iid assumption on u_t and the finiteness of the expected loss.

2. Let $\beta_1^* \neq \beta_1$. We have immediately that

$$\frac{1}{T}\sum \mathcal{L}\left(y_t - \beta_0^* - \beta_1^* x_{t-1}\right) = \frac{1}{T}\sum \mathcal{L}\left(u_t - \beta_0^* + \left(\beta_1 - \beta_1^*\right) x_{t-1}\right).$$

   But the loss function $\mathcal{L}$ is continuous and homogenous of order p, so the CMT leads to

$$\frac{1}{n_T^p T}\sum \mathcal{L}\left(y_t - \beta_0^* - \beta_1^* x_{t-1}\right) \Rightarrow \int_0^1 \mathcal{L}\left(\left(\beta_1 - \beta_1^*\right)\sigma_v X\left(s\right)\right)\mathrm{d}s;$$

   because $\mathcal{L}$ only takes nonnegative values, it follows that

$$\frac{1}{T}\sum \mathcal{L}\left(y_t - \beta_0^* - \beta_1^* x_{t-1}\right) \xrightarrow{p} \infty.$$

Since $\mathrm{E}\left(\mathcal{L}\left(u_t - \beta_0^*\right)\right)$ is finite, the target function is minimized with probability approaching 1 at β_1 as $T \to \infty$. Therefore, $\widehat{\beta}_1 \xrightarrow{p} \beta_1$ irrespective of the behavior of $\widehat{\beta}_0$ (which does not matter because of the discontinuity in limiting function).

For $\widehat{\beta}_0$, assume for simplicity that β_0 is known to belong to a compact set; then, pointwise convergence and convexity of the target function imply uniform convergence Andersen and Gill (1982, Lemma II.1) to the argmin of $\mathrm{E}\left(\mathcal{L}\left(u_t - \beta_0^*\right)\right)$. But the argmin is indeed β_0 according to its definition (10), so $\widehat{\beta}_0 \xrightarrow{p} \beta_0$ as required.

To establish the desired convergence rate, consider the sequence $\beta_1^* = \beta_1 + b/n_T$ and let w.l.o.g. $\beta_0^* = \beta_0$. Using a Taylor expansion around β_1, it follows that

$$\begin{aligned}\tfrac{1}{T}\textstyle\sum \mathcal{L}\left(y_t - \beta_0^* - \beta_1^* x_{t-1}\right) &= \tfrac{1}{T}\textstyle\sum \mathcal{L}\left(u_t - \beta_0\right) + \tfrac{b}{T}\textstyle\sum \mathcal{L}^{(1)}\left(u_t - \beta_0\right)\tfrac{x_{t-1}}{n_T} \\ &\quad + \tfrac{b^2}{T}\textstyle\sum \mathcal{L}^{(2)}\left(u_t - \beta_0 - \tfrac{b^*}{n_T}x_{t-1}\right)\left(\tfrac{x_{t-1}}{n_T}\right)^2,\end{aligned}$$

where $0 \leq b^* \leq b$. The first term on the r.h.s. converges to $\mathrm{E}\left(\mathcal{L}\left(u_t - \beta_0\right)\right)$ which is the minimum of the target function; the second converges according to Lemma 1 to zero in probability.

For the third, note that, due to the convexity of $\mathcal{L}$, $\mathcal{L}^{(2)}$ is bounded away from zero, so there exists $C > 0$ such that

$$\frac{b^2}{T}\sum \mathcal{L}^{(2)}\left(u_t - \beta_0 - \frac{b^*}{n_T}x_{t-1}\right)\left(\frac{x_{t-1}}{n_T}\right)^2 \geq \frac{Cb^2}{T}\sum\left(\frac{x_{t-1}}{n_T}\right)^2,$$

where $T^{-1}\sum\left(\frac{x_{t-1}}{n_T}\right)^2 \Rightarrow \int_0^1 X^2(s)\,\mathrm{d}s$ which is positive w.p.1. Hence, unless $b = 0$, the minimum of the target function is not achieved under $\beta_1^* = \beta_1 + b/n_T$ and $\widehat{\beta}_1$ must converge at a rate faster than n_T^{-1}, as required.

Lemma 3 *Under the definition of persistence and assumptions 1 - 3,*

$$\frac{1}{n_T T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1} \xrightarrow{p} 0,$$

as $T \to \infty$.

Proof of Lemma 3

Let $\widetilde{S}_t = \sum_{j=1}^{t}\widetilde{z}_j^{(I)}$ with $\widetilde{S}_0 \equiv 0$ such that

$$\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1} = \sum_{t=2}^{T}\left(\widetilde{S}_{t-1} - \widetilde{S}_{t-2}\right)x_{t-1} = \widetilde{S}_{T-1}x_{T-1} - \sum_{t=2}^{T-1}\widetilde{S}_{t-1}\Delta x_t. \tag{26}$$

With $z_{t-1}^{(I)} = \Delta x_{t-1}$,

$$\begin{aligned}\widetilde{S}_{T-1}x_{T-1} &= \left(\sum_{j=1}^{T-1}\left(\Delta x_j - \frac{1}{T}\sum_{t=2}^{T}\Delta x_t\right)\right)x_{T-1}\\ &= \left((x_{T-1} - x_0) - \frac{T-1}{T}(x_{T-1} - x_1)\right)x_{T-1}\\ &= O_p\left(n_T^2\right),\end{aligned}$$

under persistence. Regarding the second term on the r.h.s. in (26),

$$\sum_{t=2}^{T-1}\left(\sum_{j=1}^{t-1}\widetilde{z}_j^{(I)}\right)\Delta x_t = \sum_{t=2}^{T-1}\left(\sum_{j=1}^{t-1}\Delta x_j\right)\Delta x_t - \left(\sum_{t=2}^{T-1}\frac{t-1}{T}\Delta x_t\right)\left(\sum_{s=2}^{T}\Delta x_s\right).$$

Now $\sum_{s=2}^{T}\Delta x_s = O_p(n_T) = \sum_{t=2}^{T-1}(t-1)/T\Delta x_t$ such that

$$\left(\sum_{t=2}^{T-1}\frac{t-1}{T}\Delta x_t\right)\left(\sum_{s=2}^{T}\Delta x_s\right) = O_p\left(n_T^2\right).$$

Next, by rearranging the summation, we obtain

$$\sum_{t=2}^{T-1}\left(\sum_{j=1}^{t-1}\Delta x_j\right)\Delta x_t=\frac{1}{2}\left(\left(\sum_{t=1}^{T-1}\Delta x_t\right)^2-\sum_{t=1}^{T-1}\left(\Delta x_t\right)^2\right)=O_p\left(\max\left[n_T^2,\ T\right]\right),$$

using $\left(\sum_{t=1}^{T-1}\Delta x_t\right)^2=O_p\left(n_T^2\right)$ and $\sum_{t=2}^{T-1}\left(\Delta x_t\right)^2=O_p\left(T\right)$, where the latter result can be established by using Markov's inequality, Minkowski's inequality and the fact that $\mathbb{E}\left|\Delta x_t\right|^4$ is uniformly bounded by assumption 3. Taken together,

$$\frac{1}{n_T T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1}=O_p\left(\max\left[\frac{n_T}{T},\ \frac{1}{n_T}\right]\right),$$

the result follows since $n_T/T\to 0$ by definition 1.

Proofs of the main results

Proof of Theorem 1

We focus on the case in which x_{t-1} is persistent. The case of stationary predictors is carried out by analogous arguments and details are omitted.

Let

$$D_T=\begin{bmatrix}\sqrt{T} & 0\\ 0 & \sqrt{T}\end{bmatrix},$$

and define

$$q_T=D_T^{-1}\left(\sum_{t=2}^{T}\widetilde{\mathbf{z}}_{t-1,T}\mathcal{L}^{(1)}\left(u_t-\widehat{\beta}_0\right)\right),$$
$$Q_T=D_T^{-1}\left(\sum_{t=2}^{T}\widetilde{\mathbf{z}}_{t-1,T}\widetilde{\mathbf{z}}_{t-1,T}'\left(\mathcal{L}^{(1)}\left(u_t-\widehat{\beta}_0\right)\right)^2\right)D_T^{-1}.$$

Then $\mathcal{T}=q_T'\left(Q_T\right)^{-1}q_T$. We show $(\mathbf{i})$ that q_T converges in distribution to a normal distribution with asymptotic mean zero and asymptotic covariance matrix $\mathcal{Q}$ and $(\mathbf{ii})$ that Q_T converges in probability to $\mathcal{Q}$. The result follows then from the properties of the multivariate normal distribution.

In q_T, $\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(I)}\mathcal{L}^{(1)}\left(u_t-\widehat{\beta}_0\right)$, say, can be represented in matrix notation using the projection matrix $I_{T-1}-\iota\iota'/\left(T-1\right)$, with ι being a $(T-1)\times 1$ vector of ones. Due to the idempotency of this matrix, we can then replace $\widetilde{\mathbf{z}}_{t-1,T}$ by $\mathbf{z}_{t-1}$ without affecting the asymptotic results. Regarding $(\mathbf{i})$, note first that by the mean value theorem,

$$\mathcal{L}^{(1)}\left(u_t-\widehat{\beta}_0\right)=\mathcal{L}^{(1)}\left(u_t-\beta_0\right)-\left(\widehat{\beta}_0-\beta_0\right)\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}\right),\tag{27}$$

with $\widehat{\beta}_{0,t} = \gamma_t \beta_0 + (1-\gamma_t)\,\widehat{\beta}_0$, for some $0 \leq \gamma_t \leq 1$. Hence $\widehat{\beta}_{0,t} \xrightarrow{p} \beta_0$ uniformly over t. Moreover,

$$\sum_{t=2}^{T} \mathbf{z}_{t-1,T} \mathcal{L}^{(1)} \left(u_t - \widehat{\beta}_0 \right) = \sum_{t=2}^{T} \mathbf{z}_{t-1,T} \widetilde{u}_t - \left(\widehat{\beta}_0 - \beta_0 \right) \sum_{t=2}^{T} \mathbf{z}_{t-1,T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right),$$

and

$$q_T = \begin{bmatrix} \frac{1}{\sqrt{T}} \sum_{t=2}^{T} z_{t-1,T}^{(I)} \widetilde{u}_t - \sqrt{T} \left(\widehat{\beta}_0 - \beta_0 \right) \frac{1}{T} \sum_{t=2}^{T} z_{t-1,T}^{(I)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) \\ \frac{1}{\sqrt{T}} \sum_{t=2}^{T} z_{t-1,T}^{(II)} \widetilde{u}_t - \sqrt{T} \left(\widehat{\beta}_0 - \beta_0 \right) \frac{1}{T} \sum_{t=2}^{T} z_{t-1,T}^{(II)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) \end{bmatrix}. \tag{28}$$

Note that from (27) and the definition of $\widehat{\beta}_0$,

$$0 = \sum_{t=2}^{T} \mathcal{L}^{(1)} \left(u_t - \widehat{\beta}_0 \right) = \sum_{t=2}^{T} \mathcal{L}^{(1)} \left(u_t - \beta_0 \right) - \left(\widehat{\beta}_0 - \beta_0 \right) \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right),$$

such that

$$\sqrt{T} \left(\widehat{\beta}_0 - \beta_0 \right) = \frac{\frac{1}{\sqrt{T}} \sum_{t=2}^{T} \widetilde{u}_t}{\frac{1}{T} \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right)}. \tag{29}$$

Let

$$A_T = \begin{bmatrix} 1 & 0 & -a_{T,13} \\ 0 & 1 & -a_{T,23} \end{bmatrix}.$$

with

$$a_{T,13} = \frac{\frac{1}{T} \sum_{t=2}^{T} z_{t-1,T}^{(I)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right)}{\frac{1}{T} \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right)}, \tag{30}$$

$$a_{T,23} = \frac{\frac{1}{T} \sum_{t=2}^{T} z_{t-1,T}^{(II)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right)}{\frac{1}{T} \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right)}, \tag{31}$$

and let

$$\xi_T = \begin{bmatrix} \frac{1}{\sqrt{T}} \sum_{t=2}^{T} z_{t-1,T}^{(I)} \widetilde{u}_t \\ \frac{1}{\sqrt{T}} \sum_{t=2}^{T} z_{t-1,T}^{(II)} \widetilde{u}_t \\ \frac{1}{\sqrt{T}} \sum_{t=2}^{T} \widetilde{u}_t \end{bmatrix}.$$

Then $q_T = A_T \xi_T$. As a consequence of Assumptions 2 and 3, ξ_T is asymptotically normally distributed with asymptotic mean equal to zero and positive definite asymptotic covariance matrix V_ξ, say, $\xi_T \xrightarrow{d} \xi$, $\xi \sim \mathcal{N}(0, V_\xi)$. Furthermore, we show

$$A_T \xrightarrow{p} A, \tag{32}$$

with

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -a_{23} \end{bmatrix} = O(1),$$

where $a_{23} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=2}^{T} z_{t-1,T}^{(II)}$.

By Slutsky's theorem, we then have $q_T \xrightarrow{d} \mathcal{N}(0, \mathcal{Q})$ with $\mathcal{Q} \equiv AV_\xi A'$. Regarding (32), we verify that

$$\frac{1}{T} \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) \xrightarrow{p} \widetilde{\kappa}^{(2)}, \tag{33}$$

$$\frac{1}{T} \sum_{t=2}^{T} z_{t-1}^{(I)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) \xrightarrow{p} 0, \tag{34}$$

$$\frac{1}{T} \sum_{t=2}^{T} z_{t-1}^{(II)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) \xrightarrow{p} \widetilde{\kappa}^{(2)} \Sigma_z^{13}. \tag{35}$$

where $\Sigma_z^{13} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=2}^{T} z_{t-1,T}^{(II)}$, with Σ_z defined in assumption 3, implying

$$\begin{aligned} a_{13} &= 0 \\ a_{23} &= \Sigma_z^{13}. \end{aligned}$$

To establish (33)-(35), we use arguments similar to those in the proof of Lemma 1.6. For $p = 2$ and $p = 3$, the facts that the second derivative is piecewise constant and Lipschitz, respectively, can be employed to establish the necessary results, along the lines of the following arguments. For $p > 3$, we make repeated use of the following Taylor expansion around $u_t - \beta_0$,

$$\begin{aligned} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) = \mathcal{L}^{(2)} (u_t - \beta_0) &+ \sum_{j=3}^{p-2} \frac{1}{(j-2)!} \mathcal{L}^{(j)} (u_t - \beta_0) \left(\beta_0 - \widehat{\beta}_{0,t} \right)^{j-2} \\ &+ \frac{1}{(p-3)!} \mathcal{L}^{(p-1)} \left(u_t - \widehat{\beta}_{0,t}^{*} \right) \left(\beta_0 - \widehat{\beta}_{0,t} \right)^{p-3}. \end{aligned} \tag{36}$$

for $\widehat{\beta}_{0,t}^{*}$ between $\widehat{\beta}_{0,t}$ and β_0, and thus converging uniformly to β_0 as well, implying

$$\begin{aligned} \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) = \sum_{t=2}^{T} \mathcal{L}^{(2)} (u_t - \beta_0) &+ \sum_{t=2}^{T} \left(\sum_{j=3}^{p-2} \frac{1}{(j-2)!} \mathcal{L}^{(j)} (u_t - \beta_0) \left(\beta_0 - \widehat{\beta}_{0,t} \right)^{j-2} \right) \\ &+ \frac{1}{(p-3)!} \sum_{t=2}^{T} \mathcal{L}^{(p-1)} \left(u_t - \widehat{\beta}_{0,t}^{*} \right) \left(\beta_0 - \widehat{\beta}_{0,t} \right)^{p-3}. \end{aligned} \tag{37}$$

As a consequence of assumption 2, $\frac{1}{T} \sum_{t=2}^{T} \mathcal{L}^{(2)} (u_t - \beta_0) \xrightarrow{p} \widetilde{\kappa}^{(2)}$, where, due to convexity of $\mathcal{L}(\cdot)$ and monotonicity of expectation, $\widetilde{\kappa}^{(2)} = \mathbb{E}\left[\mathcal{L}^{(2)} (u_t - \beta_0) \right] > 0$.

Similarly, since $T^{-1}\sum_{t=2}^{T}\mathcal{L}^{(j)}\left(u_t-\widehat{\beta}_0\right)\xrightarrow{p}\widetilde{\kappa}^{(j)}$ and $\widehat{\beta}_{0,t}\xrightarrow{p}\beta_0$ uniformly,

$$\left(\beta_0-\widetilde{\beta}_0\right)^j\frac{1}{T}\sum_{t=2}^{T}\mathcal{L}^{(j)}\left(u_t-\beta_0\right)=o_p\left(1\right),$$

for $j=3,\ldots,p$. Using the Lipschitz continuity of $\mathcal{L}^{(p-1)}$,

$$\left|\mathcal{L}^{(p-1)}\left(u_t-\widehat{\beta}^*_{0,t}\right)-\mathcal{L}^{(p-1)}\left(u_t-\beta_0\right)\right|\leq C\left|\beta_0-\widehat{\beta}^*_{0,t}\right|=o_p\left(1\right),$$

and the same argument as above applies to the last term in (37). Therefore,

$$\frac{1}{T}\sum_{t=2}^{T}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}\right)\xrightarrow{p}\widetilde{\kappa}^{(2)},$$

such that (33) holds.

Turning to (34),

$$\begin{aligned}\sum_{t=2}^{T}z^{(I)}_{t-1,T}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}\right)&=\sum_{t=2}^{T}z^{(I)}_{t-1,T}\mathcal{L}^{(2)}\left(u_t-\beta_0\right)\\&+\sum_{t=2}^{T}z^{(I)}_{t-1,T}\left(\sum_{j=3}^{p-2}\frac{1}{(j-2)!}\mathcal{L}^{(j)}\left(u_t-\beta_0\right)\left(\beta_0-\widehat{\beta}_{0,t}\right)^{j-2}\right)\\&+\sum_{t=2}^{T}z^{(I)}_{t-1,T}\left(\frac{1}{(p-3)!}\mathcal{L}^{(p-1)}\left(u_t-\widehat{\beta}^*_{0,t}\right)\left(\beta_0-\widehat{\beta}_{0,t}\right)^{p-3}\right).\end{aligned}$$

First,

$$\frac{1}{T}\sum_{t=2}^{T}z^{(I)}_{t-1,T}\mathcal{L}^{(2)}\left(u_t-\beta_0\right)=\frac{1}{T}\sum_{t=2}^{T}z^{(I)}_{t-1,T}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)+\widetilde{\kappa}^{(2)}\frac{1}{T}\sum_{t=2}^{T}z^{(I)}_{t-1,T}.$$

Assumptions 2 and 3 imply that $\left\{z^{(I)}_{t-1,T}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)\right\}$ is a martingale difference (md) sequence with $\mathbb{E}\left|z^{(I)}_{t-1,T}\mathcal{L}^{(2)}\left(u_t-\beta_0\right)\right|^2=\mathbb{E}\left|\left(z^{(I)}_{t-1,T}\right)^2\right|\mathbb{E}\left|\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)\right)^2\right|<C<\infty$, which follows from assumptions 2, 3, which says that $z^{(I)}_{t-1,T}=\Delta x_{t-1}$ which in turn implies that $z^{(I)}_{t-1,T}$ is independent of $\mathcal{L}^{(2)}\left(u_t-\beta_0\right)$. Hence by a law of large numbers for md sequences (see for example White (2001), section 3.5), we have

$$\frac{1}{T}\sum_{t=2}^{T}z^{(I)}_{t-1,T}\left(\mathcal{L}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)\xrightarrow{p}0.$$

Furthermore, as a consequence of assumption 3,

$$\frac{1}{T}\sum_{t=2}^{T} z_{t-1,T}^{(I)} = \frac{n_T}{T}\left(\frac{1}{n_T}\left(x_{T-1} - x_1\right)\right) = O_p\left(\frac{n_T}{T}\right),$$

such that $\frac{1}{T}\sum_{t=2}^{T} z_{t-1,T}^{(I)}\mathcal{L}^{(2)}\left(u_t - \beta_0\right) \xrightarrow{p} 0$ by definition 1.

Similarly, for $j = 3, \ldots, (p-2)$, by adding and subtracting

$$\widetilde{\kappa}^{(j)} = \text{plim}\; T^{-1}\sum_{t=2}^{T} z_{t-1,T}^{(I)}\mathcal{L}^{(j)}\left(u_t - \beta_0\right),$$

we obtain

$$\left(\beta_0 - \widehat{\beta}_{0,t}\right)^{j-2}\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}\mathcal{L}^{(j)}\left(u_t - \beta_0\right) = o_p\left(1\right).$$

By the Lipschitz condition for $\mathcal{L}^{(p-1)}$, the same reasoning applies and we conclude that

$$\left(\beta_0 - \widehat{\beta}_{0,t}\right)^{p-3}\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}\mathcal{L}^{(p-1)}\left(u_t - \widehat{\beta}_{0,t}^{*}\right) = o_p\left(1\right).$$

Combining these arguments yields (34).

Exactly analogous arguments apply to (35). In particular,

$$\frac{1}{T}\sum_{t=1}^{T} z_{t-1,T}^{(II)}\mathcal{L}^{(2)}\left(u_t - \beta_0\right) = \frac{1}{T}\sum_{t=2}^{T} z_{t-1,T}^{(II)}\left(\mathcal{L}^{(2)}\left(u_t - \beta_0\right) - \widetilde{\kappa}^{(2)}\right) + \widetilde{\kappa}^{(2)}\frac{1}{T}\sum_{t=2}^{T} z_{t-1,T}^{(II)}$$

converges in probability to $\widetilde{\kappa}^{(2)}\Sigma_z^{13}$. Proceeding in this fashion gives (35), completing the first part of the proof.

Regarding $(\mathbf{ii})$, let $\mathcal{Q}_{ij}$ denote the (i,j) element of $\mathcal{Q}$. Here,

$$\begin{aligned}
\mathcal{Q}_{11} &= V_{11}, \\
\mathcal{Q}_{12} &= V_{12} - a_{23}V_{13}, \\
\mathcal{Q}_{22} &= V_{22} - 2a_{23}V_{23} + a_{23}^2 V_{33},
\end{aligned}$$

where V_{ij} denotes the (i,j) element of V_ξ. Here,

$$\begin{aligned}
V_{11} &= \sigma_{\widetilde{u}}^2 \lim_{T\to\infty}\frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\left[\left(z_{t-1,T}^{(I)}\right)^2\right] = \sigma_{\widetilde{u}}^2\Sigma_z^{22}, \\
V_{22} &= \sigma_{\widetilde{u}}^2 \lim_{T\to\infty}\frac{1}{T}\sum_{t=2}^{T}\left(z_{t-1,T}^{(II)}\right)^2 = \sigma_{\widetilde{u}}^2\Sigma_z^{33}, \\
V_{33} &= \sigma_{\widetilde{u}}^2,
\end{aligned}$$

which are finite under assumptions 2 and 3, again using the definition of Σ_z. Furthermore,

$$V_{12} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=2}^{T} z_{t-1,T}^{(II)} \mathbb{E}\left[z_{t-1,T}^{(I)} \widetilde{u}_t^2\right] = \sigma_{\widetilde{u}}^2 \lim_{T\to\infty} \frac{1}{T}\sum_{t=2}^{T} z_{t-1,T}^{(II)} \mathbb{E}\left[z_{t-1,T}^{(I)}\right] = 0,$$

$$V_{13} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=2}^{T} \mathbb{E}\left[z_{t-1,T}^{(I)} \widetilde{u}_t^2\right] = \sigma_{\widetilde{u}}^2 \lim_{T\to\infty} \frac{1}{T}\sum_{t=2}^{T} \mathbb{E}\left[z_{t-1,T}^{(I)}\right] = 0,$$

$$V_{23} = \sigma_{\widetilde{u}}^2 \lim_{T\to\infty} \frac{1}{T}\sum_{t=2}^{T} z_{t-1,T}^{(II)} = \sigma_{\widetilde{u}}^2 \Sigma_z^{13},$$

where $\mathbb{E}\left[z_{t-1,T}^{(I)}\right] = 0$ follows by Assumptions 2 and 3.
Using (27),

$$Q_T = \widetilde{Q}_T + R_{1T}^Q + R_{2T}^Q, \tag{38}$$

where

$$\widetilde{Q}_T = \begin{bmatrix} \frac{1}{T}\sum_{t=2}^{T}\left(\widetilde{z}_{t-1}^{(I)}\right)^2 \widetilde{u}_t^2 & \frac{1}{T}\sum_{t=2}^{T} \widetilde{z}_{t-1}^{(I)}\widetilde{z}_{t-1}^{(II)}\widetilde{u}_t^2 \\ \frac{1}{T}\sum_{t=2}^{T} \widetilde{z}_{t-1}^{(I)}\widetilde{z}_{t-1}^{(II)}\widetilde{u}_t^2 & \frac{1}{T}\sum_{t=2}^{T}\left(\widetilde{z}_{t-1}^{(II)}\right)^2 \widetilde{u}_t^2 \end{bmatrix},$$

$$R_{1T}^Q = \left(\widehat{\beta}_0 - \beta_0\right)^2 \begin{bmatrix} \frac{1}{T}\sum_{t=2}^{T}\left(\widetilde{z}_{t-1}^{(I)}\right)^2 \left(\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t}\right)\right)^2 & \frac{1}{T}\sum_{t=2}^{T} \widetilde{z}_{t-1}^{(I)}\widetilde{z}_{t-1}^{(II)} \left(\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t}\right)\right)^2 \\ \frac{1}{T}\sum_{t=2}^{T} \widetilde{z}_{t-1}^{(I)}\widetilde{z}_{t-1}^{(II)} \left(\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t}\right)\right)^2 & \frac{1}{T}\sum_{t=2}^{T}\left(\widetilde{z}_{t-1}^{(II)}\right)^2 \left(\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t}\right)\right)^2 \end{bmatrix},$$

$$R_{2T}^Q = -2\left(\widehat{\beta}_0 - \beta_0\right) \begin{bmatrix} \frac{1}{T}\sum_{t=2}^{T} \widetilde{u}_t \left(\widetilde{z}_{t-1}^{(I)}\right)^2 \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t}\right) & \frac{1}{T}\sum_{t=2}^{T} \widetilde{u}_t \widetilde{z}_{t-1}^{(I)}\widetilde{z}_{t-1}^{(II)} \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t}\right) \\ \frac{1}{T}\sum_{t=2}^{T} \widetilde{u}_t \widetilde{z}_{t-1}^{(I)}\widetilde{z}_{t-1}^{(II)} \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t}\right) & \frac{1}{T}\sum_{t=2}^{T} \widetilde{u}_t \left(\widetilde{z}_{t-1}^{(II)}\right)^2 \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t}\right) \end{bmatrix}.$$

We first verify that $\widetilde{Q}_T$ converges in probability to $\mathcal{Q}$. Notice that

$$\widetilde{z}_{t-1}^{(I)} = z_{t-1,T}^{(I)} - \widehat{a}_{13},$$
$$\widetilde{z}_{t-1}^{(II)} = z_{t-1,T}^{(II)} - \widehat{a}_{23},$$

where $\widehat{a}_{13} = T^{-1}\sum_{t=2}^{T} z_{t-1,T}^{(I)}$ and $\widehat{a}_{23} = T^{-1}\sum_{t=2}^{T} z_{t-1,T}^{(II)}$. Now

$$\frac{1}{T}\sum_{t=2}^{T}\left(z_{t-1,T}^{(I)} - \widehat{a}_{13}\right)^2 \widetilde{u}_t^2 = \frac{1}{T}\sum_{t=2}^{T}\left(z_{t-1,T}^{(I)}\right)^2 \widetilde{u}_t^2 - 2\widehat{a}_{13}\frac{1}{T}\sum_{t=2}^{T} z_{t-1,T}^{(I)}\widetilde{u}_t^2 + \left(\widehat{a}_{13}\right)^2 \frac{1}{T}\sum_{t=2}^{T}\widetilde{u}_t^2.$$

By assumptions 2 and 3, $\mathbb{E}\left|\left(z_{t-1,T}^{(I)}\right)^2 \widetilde{u}_t^2\right| < \infty$, so $\left\{\left(z_{t-1,T}^{(I)}\right)^2 \widetilde{u}_t^2 - \sigma_{\widetilde{u}}^2 \mathbb{E}\left[\left(z_{t-1,T}^{(I)}\right)^2\right]\right\}$ is a md sequence with

$$\mathbb{E}\left|\left(z_{t-1,T}^{(I)}\right)^2 \widetilde{u}_t^2\right|^2 = \mathbb{E}\left|z_{t-1,T}^{(I)}\right|^4 \mathbb{E}\left|\widetilde{u}_t\right|^4 < C < \infty,$$

which follows by construction of $z_{t-1}^{(I)}$ and from assumptions 2 and 3. Hence by a law of large numbers for md sequences

$$\frac{1}{T}\sum_{t=2}^{T}\left(z_{t-1,T}^{(I)}\right)^2\widetilde{u}_t^2=\frac{1}{T}\sum_{t=2}^{T}\left(\left(z_{t-1,T}^{(I)}\right)^2\widetilde{u}_t^2-\sigma_{\widetilde{u}}^2\mathbb{E}\left[\left(z_{t-1,T}^{(I)}\right)^2\right]\right)+\sigma_{\widetilde{u}}^2\frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\left[\left(z_{t-1,T}^{(I)}\right)^2\right]\xrightarrow{p}V_{11}.$$

Similar arguments can be invoked to show $T^{-1}\sum_{t=2}^{T}z_{t-1,T}^{(I)}\widetilde{u}_t^2\xrightarrow{p}0$ and by combining this results with $\widehat{a}_{13}\xrightarrow{p}0$ and $T^{-1}\sum_{t=2}^{T}\widetilde{u}_t^2\xrightarrow{p}\sigma_{\widetilde{u}}^2$, we have

$$\frac{1}{T}\sum_{t=2}^{T}\left(\widetilde{z}_{t-1}^{(I)}\right)^2\widetilde{u}_t^2\xrightarrow{p}\mathcal{Q}_{11}.$$

Analogous arguments apply to the other elements of $\widetilde{Q}_T$ to obtain

$$\widetilde{Q}_T\xrightarrow{p}\mathcal{Q}.$$

Consider now R_{1T}^{Q}. Following the same steps as in $(\mathbf{i})$ making use of a Taylor expansion analogous to (36), we have

$$\frac{1}{T}\sum_{t=2}^{T}\left(\widetilde{z}_{t-1}^{(I)}\right)^2\left(\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}\right)\right)^2=\frac{1}{T}\sum_{t=2}^{T}\left(\widetilde{z}_{t-1}^{(I)}\right)^2\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)\right)^2+o_p\left(1\right)=O_p\left(1\right).$$

From (29) and the preceding results it is easy to see that $\left(\widehat{\beta}_0-\beta_0\right)=o_p\left(1\right)$. Then for the $(1,1)$ element of R_{1T}^{Q} it holds that

$$\left(\widehat{\beta}_0-\beta_0\right)^2\frac{1}{T}\sum_{t=2}^{T}\left(\widetilde{z}_{t-1}^{(I)}\right)^2\left(\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}\right)\right)^2\xrightarrow{p}0.$$

Continuing in this manner,

$$R_{1T}^{Q}\xrightarrow{p}0,$$
$$R_{2T}^{Q}\xrightarrow{p}0,$$

which completes the proof of the theorem.

Proof of Theorem 2

The proof follows the steps of the proof of Theorem 1. We consider the cases of $(\mathbf{i})$ persistence and $(\mathbf{ii})$ stationarity of the predictors separately.

(**i**) Suppose x_{t-1} is persistent. By the mean value theorem

$$\mathcal{L}^{(1)}\left(y_t-\widehat{\beta}_0\right)=\mathcal{L}^{(1)}\left(u_t-\beta_0\right)-\left(\widehat{\beta}_0-\beta_0\right)\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)$$
$$+\left(\frac{b}{n_T\sqrt{T}}x_{t-1}\right)\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right),$$

for some $\widehat{\beta}_{0,t}$ between $\widehat{\beta}_0$ and β_0 and $\widehat{b}$ between b and zero. We then have

$$\sqrt{T}\left(\widehat{\beta}_0-\beta_0\right)=\frac{\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\widetilde{u}_t}{\frac{1}{T}\sum_{t=2}^{T}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)}$$
$$+\frac{\frac{b}{n_TT}\sum_{t=2}^{T}x_{t-1}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)}{\frac{1}{T}\sum_{t=2}^{T}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)}.$$

Therefore, with the scaling matrix D_T as defined in the proof of Theorem 1,

$$q_T=D_T^{-1}\left(\sum_{t=2}^{T}\widetilde{\mathbf{z}}_{t-1}\mathcal{L}^{(1)}\left(y_t-\widehat{\beta}_0\right)\right)$$
$$=\begin{bmatrix}\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(I)}\widetilde{u}_t-\sqrt{T}\left(\widehat{\beta}_0-\beta_0\right)\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(I)}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)\\ \frac{1}{\sqrt{T}}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(II)}\widetilde{u}_t-\sqrt{T}\left(\widehat{\beta}_0-\beta_0\right)\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(II)}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)\end{bmatrix}$$
$$+b\begin{bmatrix}\frac{1}{n_TT}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(I)}x_{t-1}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)\\ \frac{1}{n_TT}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(II)}x_{t-1}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)\end{bmatrix}.$$

Making use of the expression for $\sqrt{T}\left(\widehat{\beta}_0-\beta_0\right)$, we can establish the following decomposition,

$$q_T=A_T\xi_T+\Delta_T,$$

where ξ_T is defined as

$$\xi_T=\begin{bmatrix}\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(I)}\widetilde{u}_t\\ \frac{1}{\sqrt{T}}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(II)}\widetilde{u}_t\\ \frac{1}{\sqrt{T}}\sum_{t=2}^{T}\widetilde{u}_t\end{bmatrix},$$

and

$$A_T=\begin{bmatrix}1&0&-a_{T,13}\\0&1&-a_{T,23}\end{bmatrix},$$

where now

$$a_{T,13} = \frac{\frac{1}{T}\sum_{t=2}^{T} \widetilde{z}_{t-1,T}^{(I)} \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)}{\frac{1}{T}\sum_{t=2}^{T} \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)},$$

$$a_{T,23} = \frac{\frac{1}{T}\sum_{t=2}^{T} \widetilde{z}_{t-1,T}^{(II)} \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)}{\frac{1}{T}\sum_{t=2}^{T} \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)},$$

Furthermore,

$$\Delta_T = b\left(\Delta_{2,T} - \Delta_{1,T}\right).$$

where

$$\Delta_{1,T} = \begin{bmatrix} \frac{\frac{1}{T}\sum_{t=2}^{T} \widetilde{z}_{t-1,T}^{(I)} \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)}{\frac{1}{T}\sum_{t=2}^{T} \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)} \left(\frac{1}{n_T T}\sum_{t=2}^{T} x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)\right) \\ \frac{\frac{1}{T}\sum_{t=2}^{T} \widetilde{z}_{t-1,T}^{(II)} \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)}{\frac{1}{T}\sum_{t=2}^{T} \mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)} \left(\frac{1}{n_T T}\sum_{t=2}^{T} x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)\right) \end{bmatrix},$$

$$\Delta_{2,T} = \begin{bmatrix} \frac{1}{n_T T}\sum_{t=2}^{T} \widetilde{z}_{t-1,T}^{(I)} x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right) \\ + \frac{1}{n_T T}\sum_{t=2}^{T} \widetilde{z}_{t-1,T}^{(II)} x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right) \end{bmatrix}.$$

First, $A_T\xi_T$ is asymptotically normally distributed as in the proof of Theorem 1. To this end, using a Taylor expansion around $u_t - \beta_0$,

$$\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right) = \mathcal{L}^{(2)}\left(u_t - \beta_0\right) \tag{39}$$

$$+ \sum_{j=3}^{p-2} \frac{1}{(j-2)!}\mathcal{L}^{(j)}\left(u_t - \beta_0\right)\left(\beta_0 - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)^j$$

$$+ \frac{1}{(p-3)!}\mathcal{L}^{(p-1)}\left(u_t - \widehat{\beta}_{0,t}^{*} + \frac{\widehat{b}^{*}}{n_T\sqrt{T}} x_{t-1}\right)\left(\beta_0 - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)^{p-1}. \tag{40}$$

with $\widehat{b}^{*}$ between b and $\widehat{b}$. Now due to the weak convergence of x_{t-1},

$$\frac{\widehat{b}}{n_T\sqrt{T}} \sup_t \left(x_{t-1}\right) = O_p\left(T^{-1/2}\right).$$

Using similar reasoning as in the proof of lemma 2, it can be shown that $\widehat{\beta} = \beta_0 + O_p\left(T^{-1/2}\right)$, implying $\widehat{\beta}_{0,t} \xrightarrow{p} \beta_0$ uniformly, at rate $\sqrt{T}$. Therefore,

$$\sup_t \left(\beta_0 - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)^j = O_p\left(T^{-1/2}\right),$$

implying

$$\frac{1}{T}\sum_{t=2}^{T}\mathcal{L}^{(j)}\left(u_t-\beta_0\right)\left(\beta_0-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)^j$$
$$\leq \sup_t\left(\beta_0-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)^j\frac{1}{T}\sum_{t=2}^{T}\mathcal{L}^{(2)}\left(u_t-\beta_0\right)=o_p\left(1\right).$$

Using the Lipschitz continuity of $\mathcal{L}^{(p-1)}$,

$$\left|\mathcal{L}^{(p-1)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}^*}{n_T\sqrt{T}}x_{t-1}\right)-\mathcal{L}^{(p-1)}\left(u_t-\beta_0\right)\right|\leq C\left|\beta_0-\widehat{\beta}_0^*+\frac{\widehat{b}^*}{n_T\sqrt{T}}x_{t-1}\right|=o_p\left(1\right),$$

and we can employ the same reasoning for the last term in the Taylor expansion. Thus

$$\frac{1}{T}\sum_{t=2}^{T}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)\xrightarrow{p}\widetilde{\kappa}^{(2)}.$$

By similar arguments,

$$\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)\xrightarrow{p}0, \tag{41}$$
$$\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(II)}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)\xrightarrow{p}0. \tag{42}$$

Hence $q_T\xrightarrow{d}\mathcal{N}\left(0,\mathcal{Q}\right)$, with

$$Q_{11}=\sigma_{\widetilde{u}}^2\lim_{T\to\infty}\frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\left[\left(\widetilde{z}_{t-1}^{(I)}\right)^2\right], \tag{43}$$
$$Q_{22}=\sigma_{\widetilde{u}}^2\lim_{T\to\infty}\frac{1}{T}\sum_{t=2}^{T}\left(\widetilde{z}_{t-1,T}^{(II)}\right)^2, \tag{44}$$
$$Q_{12}=0,$$

where the first two limits are finite by assumption 3.

Second, we argue that $\Delta_{1,T}\xrightarrow{p}0$ and

$$\Delta_{2,T}\Rightarrow\Delta_2\equiv\widetilde{\kappa}^{(2)}\ \sigma_v\ \begin{bmatrix}0\\ \int_0^1\widetilde{Z}(s)X\left(s\right)ds\end{bmatrix}, \tag{45}$$

where $\widetilde{Z}\left(s\right)=Z\left(s\right)-\int_0^1 Z\left(r\right)dr$ with $Z\left(s\right)=\sin\left(s\pi/2\right)$, say. The proof of the theorem follows then by combining (44), $\Delta_{1,T}\xrightarrow{p}0$, and (45) and the properties of the non-central χ^2 distribution.

Regarding $\Delta_{1,T}$, notice that

$$\frac{1}{n_T T}\sum_{t=2}^{T} x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right) = \frac{1}{n_T T}\sum_{t=2}^{T} x_{t-1}\mathcal{L}^{(2)}\left(u_t - \beta_0\right) + o_p\left(1\right),$$

such that the result follows by applying lemma 1, item 3, and combining this result with (41) and (42).

Finally, for $\Delta_{2,T}$, consider the first component

$$\frac{1}{n_T T}\sum_{t=2}^{T} \widetilde{z}^{(I)}_{t-1,T}x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right) = \frac{1}{n_T T}\sum_{t=2}^{T} z^{(I)}_{t-1,T}x_{t-1}\left(\mathcal{L}^{(2)}\left(u_t - \beta_0\right) - \widetilde{\kappa}^{(2)}\right)$$
$$-\left(\frac{n_T}{T}\frac{1}{n_T}\sum_{t=2}^{T} z^{(I)}_{t-1}\right)\left(\frac{1}{n_T T}\sum_{t=2}^{T} x_{t-1}\left(\mathcal{L}^{(2)}\left(u_t - \beta_0\right) - \widetilde{\kappa}^{(2)}\right)\right) + \widetilde{\kappa}^{(2)}\frac{1}{n_T T}\sum_{t=2}^{T}\widetilde{z}^{(I)}_{t-1,T}x_{t-1} + o_p\left(1\right),$$

where we made use of (39). The first term on the r.h.s. converges to zero in probability using the fact that $\left\{z^{(I)}_{t-1}n_T^{-1}x_{t-1}\left(\mathcal{L}^{(2)}\left(u_t - \beta_0\right) - \widetilde{\kappa}^{(2)}\right)\right\}$ is a md sequence with uniformly bounded second moments. Similarly, $1/\left(n_T T\right)\sum_{t=2}^{T} x_{t-1}\left(\mathcal{L}^{(2)}\left(u_t - \beta_0\right) - \widetilde{\kappa}^{(2)}\right)$ converges in probability to zero. Moreover, $1/n_T\sum_{t=2}^{T} z^{(I)}_{t-1} = 1/n_T\sum_{t=2}^{T}\Delta x_{t-1} = O_p\left(1\right)$, so the second term vanishes using $n_T/T \to 0$ by definition 1. Finally, the third term converges to zero in probability by lemma 3.

We can invoke analogous arguments for the second component of $\Delta_{2,T}$:

$$\frac{1}{n_T T}\sum_{t=2}^{T} \widetilde{z}^{(II)}_{t-1,T}x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right) = \widetilde{\kappa}^{(2)}\frac{1}{n_T T}\sum_{t=2}^{T}\widetilde{z}^{(II)}_{t-1,T}x_{t-1} + o_p\left(1\right).$$

The convergence of $\Delta_{2,T}$ follows from Lemma 3.

(**ii**) Let us now consider the stationary case. Proceeding analogously as in (**i**), we have

$$\Delta_{1,T} = \begin{bmatrix} \frac{\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}^{(I)}_{t-1,T}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)}{\frac{1}{T}\sum_{t=2}^{T}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)}\left(\frac{1}{T}\sum_{t=2}^{T} x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)\right) \\ \frac{\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}^{(II)}_{t-1,T}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)}{\frac{1}{T}\sum_{t=2}^{T}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)}\left(\frac{1}{T}\sum_{t=2}^{T} x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)\right) \end{bmatrix},$$

$$\Delta_{2,T} = \begin{bmatrix} \frac{1}{T}\sum_{t=2}^{T}\widetilde{z}^{(I)}_{t-1,T}x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right) \\ \frac{1}{T}\sum_{t=2}^{T}\widetilde{z}^{(II)}_{t-1,T}x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right) \end{bmatrix}.$$

First, $\Delta_{1,T} \xrightarrow{p} 0$. To see this, note that $1/T\sum_{t=2}^{T} x_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)$ converges to zero in probability while the same holds for

$$\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}^{(II)}_{t-1}\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right) = \frac{1}{T}\sum_{t=2}^{T}\widetilde{z}^{(II)}_{t-1}\left(\mathcal{L}^{(2)}\left(u_t - \beta_0\right) - \widetilde{\kappa}^{(2)}\right) + o_p\left(1\right).$$

Moreover, the denominator in both elements of $\Delta_{1,T}$ converges in probability to $\widetilde{\kappa}^{(2)}$ and

$$\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)=\frac{1}{T}\sum_{t=2}^{T}z_{t-1}^{(I)}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)$$
$$-\left(\frac{1}{T}\sum_{t=2}^{T}z_{t-1}^{(I)}\right)\left(\frac{1}{T}\sum_{t=2}^{T}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)\right)+o_p\left(1\right).$$

The first term on the r.h.s. converges in probability to zero by the law of large numbers for md sequences while $1/T\sum_{t=2}^{T}z_{t-1}^{(I)}=1/T\left(x_{T-1}-x_1\right)$ with $\mathbb{E}\left[x_t/T\right]=x_0/T=O\left(1/T\right)$ and $Var\left[x_t/T\right]=\sigma_v^2/T^2\sum_{j=0}^{t}\psi_{j,T}^2=O\left(1/T^2\right)$ using $\sum_{j=0}^{\infty}\psi_{j,T}^2<C<\infty$ under stationarity. Hence x_t/T converges in probability to zero. Since $1/T\sum_{t=2}^{T}\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}$ converges to zero as well, $\Delta_{1,T}\xrightarrow{p}0$.

Regarding $\Delta_{2,T}$, notice first

$$\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(II)}x_{t-1}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)=\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(II)}x_{t-1}\left(\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_0\right)-\widetilde{\kappa}^{(2)}\right)$$
$$+\widetilde{\kappa}^{(2)}\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1,T}^{(II)}x_{t-1}+o_p\left(1\right).$$

Using the fact that $\widetilde{z}_{t-1}^{(II)}$ is deterministic, the first term converges in probability to zero using the law of large numbers for md sequences. The second term equals $1/T\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(II)}\left(x_{t-1}-x_0\right)$, which converges in probability to zero as $T\to\infty$. Hence the second component of $\Delta_{2,T}$ converges to zero in probability.

Similarly,

$$\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)=\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1}\left(\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_0\right)-\widetilde{\kappa}^{(2)}\right)$$
$$+\widetilde{\kappa}^{(2)}\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1}+o_p\left(1\right),\tag{46}$$

and the first term can be further decomposed into

$$\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)=\frac{1}{T}\sum_{t=2}^{T}z_{t-1}^{(I)}x_{t-1}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)$$
$$-\left(\frac{1}{T}\sum_{t=2}^{T}z_{t-1}\right)\left(\sum_{t=2}^{T}x_{t-1}\left(\mathcal{L}^{(2)}\left(u_t-\beta_0\right)-\widetilde{\kappa}^{(2)}\right)\right)\xrightarrow{p}0,$$

using the law of large numbers for md sequences and the fact that $1/T\sum_{t=2}^{T}z_{t-1}^{(I)}$ vanishes as

$T \to \infty$. Finally, turning to the second term in (46),

$$\begin{aligned}\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1} &= \frac{1}{T}\sum_{t=2}^{T}\Delta x_{t-1}x_{t-1} - \left(\frac{1}{T}\sum_{t=2}^{T}x_{t-1}\right)\left(\frac{1}{T}\sum_{t=2}^{T}\Delta x_{t-1}\right) \\ &= \frac{1}{T}\sum_{t=2}^{T}x_{t-1}^2 - \frac{1}{T}\sum_{t=2}^{T}x_{t-2}x_{t-1} - \left(\frac{1}{T}\sum_{t=2}^{T}x_{t-1}\right)\left(\frac{1}{T}\sum_{t=2}^{T}\Delta x_{t-1}\right).\end{aligned}$$

Now $1/T\sum_{t=2}^{T}x_{t-1} \xrightarrow{p} x_0$ and $1/T\sum_{t=2}^{T}\Delta x_{t-1} = 1/T\left(x_{T-1} - x_1\right)$ which converges in mean square to zero as argued above. Hence $\left(1/T\sum_{t=2}^{T}x_{t-1}\right)\left(1/T\sum_{t=2}^{T}\Delta x_{t-1}\right) \xrightarrow{p} 0$. Moreover,

$$\frac{1}{T}\sum_{t=2}^{T}\Delta x_{t-1}x_{t-1} = \frac{1}{T}\sum_{t=2}^{T}x_{t-1}^2 - \frac{1}{T}\sum_{t=2}^{T}x_{t-1}x_{t-2},$$

and, with $\psi_{0,T} = 1$, $1/T\sum_{t=2}^{T}x_{t-1}^2 = 1/T\sum_{t=2}^{T}\left(x_0 + \sum_{j=0}^{t}\psi_{j,T}v_{t-j}\right)^2$ which converges in probability to $x_0^2 + \sigma_v^2\lim_{T\to\infty}\sum_{j=0}^{\infty}\psi_{j,t}^2$. An analogous argument applies to $1/T\sum_{t=2}^{T}x_{t-1}x_{t-2}$ to conclude

$$\frac{1}{T}\sum_{t=2}^{T}\Delta x_{t-1}x_{t-1} \xrightarrow{p} \sigma_v^2\sum_{j=0}^{\infty}\left(\psi_{j,T}^2 - \psi_{j,T}\psi_{j+1,T}\right).$$

The result follows by combining this result with (43) and the properties of the non-central χ^2 distribution.

Proof of Theorem 3

Take the Taylor expansion of the first-order conditions around $(\beta_0, \beta_1)'$ and evaluate at $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)'$.

$$\left.\begin{pmatrix} \frac{\partial}{\partial\beta_0^*}\left(\sum\mathcal{L}\left(y_t - \beta_1^*x_{t-1} - \beta_0^*\right)\right) \\ \frac{\partial}{\partial\beta_1^*}\left(\sum\mathcal{L}\left(y_t - \beta_1^*x_{t-1} - \beta_0^*\right)\right)\end{pmatrix}\right|_{\substack{\beta_0^*=\widehat{\beta}_0\\ \beta_1^*=\widehat{\beta}_1}} = \left.\begin{pmatrix} \frac{\partial}{\partial\beta_0^*}\left(\sum\mathcal{L}\left(y_t - \beta_1^*x_{t-1} - \beta_0^*\right)\right) \\ \frac{\partial}{\partial\beta_1^*}\left(\sum\mathcal{L}\left(y_t - \beta_1^*x_{t-1} - \beta_0^*\right)\right)\end{pmatrix}\right|_{\substack{\beta_0^*=\beta_0\\ \beta_1^*=\beta_1}} +$$

$$\left.\begin{pmatrix} \frac{\partial^2}{\partial\left(\beta_0^*\right)^2}\left(\sum\mathcal{L}\left(y_t - \beta_1^*x_{t-1} - \beta_0^*\right)\right) & \frac{\partial^2}{\partial\beta_1^*\partial\beta_0^*}\left(\sum\mathcal{L}\left(y_t - \beta_1^*x_{t-1} - \beta_0^*\right)\right) \\ \frac{\partial^2}{\partial\beta_0^*\partial\beta_1^*}\left(\sum\mathcal{L}\left(y_t - \beta_1^*x_{t-1} - \beta_0^*\right)\right) & \frac{\partial^2}{\partial\left(\beta_1^*\right)^2}\left(\sum\mathcal{L}\left(y_t - \beta_1^*x_{t-1} - \beta_0^*\right)\right)\end{pmatrix}\right|_{\substack{\beta_0^*=\widetilde{\beta}_0\\ \beta_1^*=\widetilde{\beta}_1}} \begin{pmatrix} \widehat{\beta}_0 - \beta_0 \\ \widehat{\beta}_1 - \beta_1\end{pmatrix},$$

where $\widetilde{\beta}_0$ and $\widetilde{\beta}_1$ lie between β_0 and $\widehat{\beta}_0$, and β_1 and $\widehat{\beta}_1$, respectively. Evaluated at $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)'$, the gradient is 0, so with $\widetilde{u}_t = \mathcal{L}^{(1)}(u_t - \beta_0)$

$$\begin{pmatrix} \sum \widetilde{u}_t \\ \sum x_{t-1}\widetilde{u}_t \end{pmatrix} = - \begin{pmatrix} \sum \mathcal{L}^{(2)}\left(y_t - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_{t-1}\right) & \sum x_{t-1}\mathcal{L}^{(2)}\left(y_t - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_{t-1}\right) \\ \sum x_{t-1}\mathcal{L}^{(2)}\left(y_t - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_{t-1}\right) & \sum x_{t-1}^2\mathcal{L}^{(2)}\left(y_t - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_{t-1}\right) \end{pmatrix} \begin{pmatrix} \widehat{\beta}_0 - \beta_0 \\ \widehat{\beta}_1 - \beta_1 \end{pmatrix},$$

Note that, since $\widehat{\beta}_1 - \beta_1$ is $o_p\left(n_T^{-1}\right)$, so must be $\widetilde{\beta}_1 - \beta_1$; also $\widetilde{\beta}_0 - \beta_0 = o_p(1)$.
Using Lemma 1 item 5, it follows that

$$\begin{aligned} &- \begin{pmatrix} \frac{1}{T}\sum \mathcal{L}^{(2)}\left(y_t - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_{t-1}\right) & \frac{1}{n_T T}\sum x_{t-1}\mathcal{L}^{(2)}\left(y_t - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_{t-1}\right) \\ \frac{1}{n_T T}\sum x_{t-1}\mathcal{L}^{(2)}\left(y_t - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_{t-1}\right) & \frac{1}{n_T^2 T}\sum x_{t-1}^2\mathcal{L}^{(2)}\left(y_t - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_{t-1}\right) \end{pmatrix}^{-1} \\ &= - \begin{pmatrix} \frac{1}{T}\sum \mathcal{L}^{(2)}(u_t - \beta_0) & \frac{1}{n_T T}\sum x_{t-1}\mathcal{L}^{(2)}(u_t - \beta_0) \\ \frac{1}{n_T T}\sum x_{t-1}\mathcal{L}^{(2)}(u_t - \beta_0) & \frac{1}{n_T^2 T}\sum x_{t-1}^2\mathcal{L}^{(2)}(u_t - \beta_0) \end{pmatrix}^{-1} + o_p(1), \end{aligned}$$

where the matrix on the r.h.s. is nonsingular with probability approaching 1. Therefore, up to an $o_p(1)$ term, we have

$$\begin{pmatrix} \sqrt{T}\left(\widehat{\beta}_0 - \beta_0\right) \\ n_T\sqrt{T}\left(\widehat{\beta}_1 - \beta_1\right) \end{pmatrix} = - \begin{pmatrix} \frac{1}{T}\sum \mathcal{L}^{(2)}(u_t - \beta_0) & \frac{1}{n_T T}\sum x_{t-1}\mathcal{L}^{(2)}(u_t - \beta_0) \\ \frac{1}{n_T T}\sum x_{t-1}\mathcal{L}^{(2)}(u_t - \beta_0) & \frac{1}{n_T^2 T}\sum x_{t-1}^2\mathcal{L}^{(2)}(u_t - \beta_0) \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sqrt{T}}\sum \widetilde{u}_t \\ \frac{1}{n_T\sqrt{T}}\sum x_{t-1}\widetilde{u}_t \end{pmatrix},$$

or

$$n_T\sqrt{T}\left(\widehat{\beta}_1 - \beta_1\right) = \frac{\frac{1}{n_T T^{1.5}}A_{1T}}{\frac{1}{n_T^2 T^2}B_{1T}} + o_p(1),$$

with

$$\begin{aligned} A_{1T} &= \sum \mathcal{L}^{(2)}(u_t - \beta_0)\sum x_{t-1}\widetilde{u}_t - \sum x_{t-1}\mathcal{L}^{(2)}(u_t - \beta_0)\sum \widetilde{u}_t \\ B_{1T} &= \sum \mathcal{L}^{(2)}(u_t - \beta_0)\sum x_{t-1}^2\mathcal{L}^{(2)}(u_t - \beta_0) - \left(\sum x_{t-1}\mathcal{L}^{(2)}(u_t - \beta_0)\right)^2, \end{aligned}$$

and

$$\sqrt{T}\left(\widehat{\beta}_0 - \beta_0\right) = \frac{\frac{1}{n_T^2 T^{1.5}}A_{0T}}{\frac{1}{n_T^2 T^2}B_{1T}} + o_p(1),$$

with

$$A_{0T} = \sum x_{t-1}^2\mathcal{L}^{(2)}(u_t - \beta_0)\sum \widetilde{u}_t - \sum x_{t-1}\mathcal{L}^{(2)}(u_t - \beta_0)\sum x_{t-1}\widetilde{u}_t,$$

leading with Lemma 1 to the desired result.

Proof of Theorem 4

Using standard regression algebra, the standard error of $\widehat{\beta}_1$ is easily checked to be given by

$$s.e.\left(\widehat{\beta}_1\right) = \sqrt{M_{1T}B_T^{-2}},$$

where

$$M_{1T} = \left(\sum \mathcal{L}^{(2)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right)\right)^2 \sum x_{t-1}^2 \left(\mathcal{L}^{(1)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right)\right)^2$$
$$+ \left(\sum x_{t-1}\mathcal{L}^{(2)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right)\right)^2 \sum \left(\mathcal{L}^{(1)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right)\right)^2 - 2 \cdot$$
$$\sum \mathcal{L}^{(2)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right) \sum x_{t-1}\mathcal{L}^{(2)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right) \sum x_{t-1}\left(\mathcal{L}^{(1)}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t-1}\right)\right)^2,$$

such that, using Lemma 1 item 5 as before,

$$\begin{aligned} M_{1T} &= \left(\sum \mathcal{L}^{(2)}\left(u_t - \beta_0\right)\right)^2 \sum x_{t-1}^2 \widetilde{u}_t^2 + \left(\sum x_{t-1}\mathcal{L}^{(2)}\left(u_t - \beta_0\right)\right)^2 \sum \widetilde{u}_t^2 \\ &\quad -2\sum \mathcal{L}^{(2)}\left(u_t - \beta_0\right) \sum x_{t-1}\mathcal{L}^{(2)}\left(u_t - \beta_0\right) \sum x_{t-1}\widetilde{u}_t^2 + o_p\left(n_T^2 T^3\right). \end{aligned}$$

Thus,

$$t_{\beta_1} = \frac{\frac{1}{n_T T^{1.5}} A_{1T}}{\sqrt{\frac{1}{n_T^2 T^3} M_{1T}}} + o_p\left(1\right),$$

and the result follows with lemma 1.

References

Aiolfi, M., M. Rodrigues, and A. Timmermann (2010). Understanding enalysts' earnings expectations: biases, nonlinearities and predictability. *Journal of Financial Econometrics 8*(3), 305–334.

Amemyia, T. (1985). *Advanced Econometrics*. Harvard University Press.

Andersen, P. K. and R. D. Gill (1982). Cox's Regression Model for Counting Processes: A large sample study. *The Annals of Statistics 10*(4), 1100–1120.

Aretz, K., S. Bartram, and P. Pope (2011). Asymmetric loss functions and the rationality of expected stock returns. *International Journal of Forecasting 27*(2), 413–437.

Artis, M. and M. Marcellino (2001). Fiscal forecasting: the track record of the IMF, OECD and EC. *The Econometrics Journal 4*(1), 20–36.

Batchelor, R. and D. Peel (1998). Rationality testing under asymmetric loss. *Economics Letters 61*(1), 49–54.

Breitung, J. and M. Demetrescu (2013). Instrumental Variable and Variable Addition Based Inference in Predictive Regressions. *Journal of Econometrics*, forthcoming.

Campbell, B. and J. M. Dufour (1995). Exact Nonparametric Orthogonality and Random Walk Tests. *The Review of Economics and Statistics 77*(1), 1–16.

Campbell, J. Y. and M. Yogo (2006). Efficient Tests of Stock Return Predictability. *Journal of Financial Economics 81*(1), 27–60.

Capistrán, C. (2008). Bias in Federal Reserve inflation forecasts: Is the Federal Reserve irrational or just cautious? *Journal of Monetary Economics 55*(8), 1415–1427.

Cavanagh, C. L., G. Elliott, and J. H. Stock (1995). Inference in Models with Nearly Integrated Regressors. *Econometric Theory 11*(5), 1131–1147.

Christodoulakis, G. and E. Mamatzakis (2013). Behavioural asymmetries in the G7 foreign exchange market. *International Review of Financial Analysis 29*, 261–270.

Christodoulakis, G. A. and E. C. Mamatzakis (2008). An Assessment of the EU Growth Forecasts under Asymmetric Preferences. *Journal of Forecasting 27*(6), 483–492.

Christodoulakis, G. A. and E. C. Mamatzakis (2009). Assessing the Prudence of Economic Forecasts in the EU. *Journal of Applied Econometrics 24*(4), 583–606.

Christoffersen, P. F. and F. X. Diebold (1997). Optimal prediction under asymmetric loss. *Econometric Theory 13*(06), 808–817.

Clatworthy, M., D. Peel, and P. Pope (2012). Are analysts' loss functions asymmetric? *Journal of Forecasting 31*(8), 736–756.

Davidson, J. and P. Sibbertsen (2005). Generating Schemes for Long Memory Processes: Regimes, aggregation and linearity. *Journal of Econometrics 128*(2), 253–282.

Dolado, J. J. and H. Lütkepohl (1996). Making Wald Tests Work for Cointegrated VAR Systems. *Econometric Reviews 15*(4), 369–386.

Elliott, G., I. Komunjer, and A. Timmermann (2005). Estimation and Testing of Forecast Rationality Under Flexible Loss. *Review of Economic Studies 72*(4), 1107–1125.

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.

Engel, C. (1996). The forward discount anomaly and the risk premium: a survey of recent evidence. *Journal of Empirial Finance 3*(2), 123–192.

Fama, E. (1984). Forward and spot exchange rates. *Journal of Monetary Economics 14*(3), 319–338.

Geweke, J. and E. Feige (1979). Some joint tests of the efficiency of markets for forward foreign exchange. *The Review of Economics and Statistics 61*(3), 334–341.

Granger, C. (1969). Prediction with a Generalized Cost of Error Function. *Operational Research Quarterly 20*(2), 199–207.

Hansen, L. and R. Hodrick (1980). Forward exchange rates as optimal predictors of future spot rates: an econometric analysis. *Journal of Political Economy 88*(5), 829–853.

Hentschel, L. (1995). All in the family: nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics 39*(1), 71–104.

Hjalmarsson, E. (2010). Predicting Global Stock Returns. *Journal of Financial and Quantitative Analysis 45*(1), 49–80.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Jansson, M. and M. J. Moreira (2006). Optimal Inference in Regression Models with Nearly Integrated Regressors. *Econometrica 74*(3), 681–714.

Komunjer, I. and M. Owyang (2012). Multivariate forecast evaluation and rationality testing. *The Review of Economics and Statistics 94*(4), 1066–1080.

Kurtz, T. G. and P. Protter (1991). Weak Limit Theorems for Stochastic Integrals and Stochastic Differential Equations. *Annals of Probability 19*(3), 1035–1070.

Lee, J. (2012). Predictive Quantile Regressions with persistent covariates. Working paper, Department of Economics, Yale University.

Lewis, K. (1995). Puzzles in international financial markets. In G. Grossman and K. Rogoff (Eds.), *Handbook of International Economics*, Volume 3, Chapter 37, pp. 1913–1971. Elsevier.

Liu, W. and A. Maynard (2005). Testing forward rate unbiasedness allowing for persistent regressors. *Journal of Empirical Finance 12*(5), 613–628.

Lucas, A. (1995). Unit root tests based on M estimators. *Econometric Theory 11*(2), 331–331.

Maynard, A. and P. C. B. Phillips (2001). Rethinking an Old Empirical Puzzle: Econometric evidence on the forward discount anomaly. *Journal of Applied Econometrics 16*(6), 671–708.

Maynard, A., K. Shimotsu, and Y. Wang (2011). Inference in Predictive Quantile Regressions. Working paper, Department of Economics, University of Guelph.

McDonald, J. B. and W. K. Newey (1988). Partially Adaptive Estimation of Regression Models via the Generalized T Distribution. *Econometric Theory 4*(3), 428–457.

Müller, U. K. and M. W. Watson (2008). Testing Models of Low-Frequency Variability. *Econometrica 76*(5), 979–1016.

Nelson, D. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica 59*(2), 347–370.

Patton, A. and A. Timmermann (2007). Testing forecast optimality under unknown loss. *Journal of the American Statistical Association 102*(480), 1172–1184.

Phillips, P. C. B. (1991). A Shortcut to LAD Estimator Asymptotics. *Econometric Theory 7*(04), 450–463.

Phillips, P. C. B. (1998). New Tools for Understanding Spurious Regressions. *Econometrica 66*(6), 1299–1325.

Phillips, P. C. B. and J. H. Lee (2013). Predictive regression under various degrees of persistence and robust long-horizon regression. *Journal of Econometrics 177*(2), 250–264.

Phillips, P. C. B. and T. Magdalinos (2007). Limit theory for moderate deviations from a unit root. *Journal of Econometrics 136*(1), 115–130.

Pierdzioch, C., J. Rülke, and G. Stadtmann (2012a). Exchange rate forecasts and asymmetric loss: empirical evidence for the yen/dollar exchange rate. *Applied Economics Letters 19*, 1759–1763.

Pierdzioch, C., J. Rülke, and G. Stadtmann (2012b). On the loss function of the Bank of Canada: A note. *Economics Letters 115*(2), 155–159.

Stambaugh, R. F. (1999). Predictive Regressions. *Journal of Financial Economics 54*(3), 375–421.

Toda, H. Y. and T. Yamamoto (1995). Statistical Inference in Vector Autoregressions with Possibly Integrated Processes. *Journal of Econometrics 66*(1-2), 225–250.

Weiss, A. A. (1996). Estimating Time Series Models Using the Relevant Cost Function. *Journal of Applied Econometrics 11*(5), 539–560.

Welch, I. and A. Goyal (2008). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies 21*(4), 1455–1508.