Three Essays in Econometrics

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften
durch die
Rechts- und Staatswissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität
Bonn

vorgelegt von
Christoph Roling
aus Coesfeld

Bonn 2014

Dekan:	Prof. Dr. Klaus Sandmann
Erstreferent:	Prof. Dr. Jörg Breitung
Zweitreferent:	Prof. Dr. Matei Demetrescu

Tag der mündlichen Prüfung: 04.02.2014

Für meine Eltern und meinen Bruder

Acknowledgements

A number of people enabled me to write this dissertation. First and foremost I am indebted to Jörg Breitung for his guidance and continued support. His insightful comments and questions sparked my interest in econometric research, helped me to enhance my understanding of the research topics and motivated me to keep improving my work.
Furthermore, I am grateful to Matei Demetrescu and Nazarii Salish for examining some of these subjects with me, which turned out to be a fruitful learning experience for me.
I would like to thank Norbert Christopeit and Christian Pigorsch, who are members of the institute for econometrics in Bonn, for econometric and non-econometric discussions and helpful advice.
I would also like to express my appreciation for Heide Baumung, Silke Kinzig and Pamela Mertens who overcame administrative obstacles, allowing me, among other things, to visit conferences in order to present excerpts of this dissertation and to be introduced into the research community in econometrics.
Moreover, I am thankful for the opportunity to become a member of the Bonn Graduate School of Economics chaired by Urs Schweizer and for financial support from the Deutsche Forschungsgemeinschaft (DFG).
My fellow doctoral students Carsten Dahremöller, Gerrit Frackenpohl, Fabian Kosse, Stephan Luck, and Ronald Rühmkorf deserve special mention. I benefited from conversations with them allowing me to glimpse at other fields of economic research or to simply have a good time.
Last but not least I would like to thank my family for their love and support, and my friends for enriching my life. You know who you are.

Contents

List of Figures

List of Tables

Introduction

Economists examine extensive cross sectional and time series data. This information allows them to study economic decision making of households, firms and countries over time or to produce forecasts of financial time series to support portfolio allocation and risk management. When analyzing and interpreting these datasets, the linear regression model continues to be fundamental to sound empirical work. The three chapters in this thesis contribute to solving several econometric issues in linear regression analysis. In the first two chapters, a potentially large number of regression parameters arises in two distinct econometric frameworks. Estimating these parameters to produce accurate forecasts of an economic variable of interest is the objective in the first chapter. The inferential methods presented in the second chapter enable the researcher to decide whether estimating a high-dimensional parameter vector is appropriate or whether a more parsimonious regression model applies. In contrast, chapter 3 shifts attention to the predictability of economic time series under a general statistical loss function.

In particular, chapter 1 examines forecasting regressions that employ many predictors, leading to the task of estimating a large number of parameters in a linear regression. Here, many regressors arise naturally due to a frequency mismatch between the series of interest and the series that is considered to have predictive power. These mixed frequency regression models arise often in macroeconomics and finance, if, for instance, quarterly g.d.p. or the monthly volatility of a return index is forecasted with daily observations of macroeconomic leading indicators or (intra-)daily observations of financial variables. A new estimation procedure for these mixed data sampling (MIDAS) regression models is proposed. The estimator is a modified ordinary least squares (OLS) estimator which assumes that the weights assigned to high-frequency regressors are a smooth function and complements the least squares criterion by a smoothness penalty, resulting in a penalized least squares (PLS) estimator. The estimation method does not rely on a particular parametric specification of the weighting function, but depends on a smoothing parameter. Several methods are presented to choose this parameter including a variant of the Akaike information criterion (AIC). A simulation study is conducted to evaluate the esti-

mation accuracy as measured by the mean squared error of the modified OLS estimator and the parametric MIDAS approach, which requires estimation by non-linear least squares (NLS). The simulation results illustrate in which cases the PLS estimator produces more accurate estimates than the parametric NLS estimator, and in which cases the parametric approach performs better. The results show that the PLS estimator is flexible alternative method to estimate MIDAS regression models. These MIDAS approaches are then employed to forecast volatility of the German stock index (DAX). In addition to the mixed frequency models, the GARCH(1,1) model is used as a benchmark. Using current and lagged absolute returns as predictors, MIDAS-PLS provides more precise forecasts than MIDAS-NLS or the GARCH(1,1) model over biweekly or monthly forecasting horizons. In a companion paper, the PLS estimator is used to forecast the monthly German inflation rate, see Breitung et al. (2013).

In chapter 2, which is joint work with Jörg Breitung and Nazarii Salish, the linear panel data model is studied, in which observations are available both in the cross section and in the time series dimension. When estimating a panel regression, it must be decided whether the economic relationship of interest is taken to be homogenous, such that the same parameter vector applies to all cross sectional units, or whether a heterogeneous empirical model is more appropriate, in which the regression parameters differ between cross sectional units. In a classical panel data setup, in which the cross section dimension is large relative to the time series dimension, modelling regression parameters to be indiviudal-specific introduces a large number of parameters to be estimated in the panel even if only a few explanatory variables are studied for each cross section unit individually. In this chapter, a statistical test is proposed to determine whether regression parameters are indiviudal-specific or common to all units in the cross section. Answering this question is important as a preparatory step in panel data analysis to select a parsimonious model if possible, and to choose the subsequent estimation procedure accordingly.

To this end, the Lagrange Multiplier (LM) principle is employed to test parameter homogeneity across cross-section units in panel data models. The test can be seen as a generalization of the Breusch-Pagan test against random individual effects to all regression coefficients. While the original test procedure assumes a likelihood framework under normality, several useful variants of the LM test are presented to allow for non-normality and heteroskedasticity. Moreover, the tests can be conveniently computed via simple artificial regressions. The limiting distribution of the LM test is derived and it is shown that if the errors are not normally distributed, the

original LM test is asymptotically valid if the number of time periods tends to infinity. A simple modification of the score statistic yields an LM test that is robust to normality if the time dimension is fixed. A further adjustment provides a heteroskedasticity-robust version of the LM test. The local power of these tests LM tests and the delta statistic proposed by Pesaran and Yamagata (2008) is then compared. The results of our Monte Carlo experiments suggest that the LM-type test can be substantially more powerful, in particular, when the number of time periods is small.

Chapter 3, written in collaborative work with Matei Demetrescu, investigates the predictive regression model under a general statistical loss function, including mean squared error (MSE) loss as a special case. In this predictive regression, the coefficient of a predictor is tested for significance. While for stationary predictors this task does not pose difficulties, non-standard limiting distributions of standard inference methods arise once the regressors are endogenous, such that there is contemporaneous correlation between shocks of the regressor and the dependent variable, and persistent, so that the predictor is reverting very slowly to its long-run mean, if at all. With a more general loss function beyond squared error loss, endogeneity is loss-function specific. Thus, no endogeneity under OLS 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. An overidentified instrumental variable based test is proposed, using particular trending functions and transformations of the predictor as instruments. The test statistic is analogous to the Anderson-Rubin statistic but takes the relevant loss into account, and follows a chi-squared distribution asymptotically, irrespective of the degree of persistence of the predictor. The forward premium puzzle is then reexamined by providing evidence for deviations from MSE loss and by conducting robust inference of the rational expectations hypothesis. The analysis provides little evidence for failure of the rational expectations hypothesis, in contrast to early empirical tests in this literature.

Chapter 1

Forecasting volatility with penalized MIDAS regressions

1.1 Introduction

Increased availability of financial and macroeconomic data enables researchers to forecast economic variables with a large number of potential predictors. In particular, predictors may be observed much more frequently than an economic variable of interest such as gross domestic product, industrial production or inflation. For instance, between two quarters in a given year, daily observations of a financial time series can be used to construct forecasts of the low-frequency variable of interest. Two tasks arise in this context. First, a forecasting model is needed to accommodate variables that are naturally measured at different frequencies. Second, the potential gains of these mixed frequency models need to be evaluated, usually relative to forecasts that do not incorporate available high-frequency data.

To deal with mixed frequencies, one can first average high-frequency observations and then produce forecasts using the implied equal-frequency framework. In this way, high-frequeny observations are equally weighted. An alternative approach allows weights to differ between intraperiod observations. For example, when using daily financial data to forecast monthly interest rates, larger weights may be assigned to daily observations that are close to the end of the current month which may reflect that market participants continuously update their expectations as new information becomes available; see Mishkin (1981), for example.

To consider these different approaches in a simple forecasting model, let

$$y_{t+h} = \alpha_0 + \sum_{p=0}^{P-1} \beta_p x_{p,t} + u_{t+h},$$

for $t = 1, \ldots, T$, where y_{t+h} is the variable to be forecasted at horizon h, and is sampled at a different frequency than the predictor x_t. The disturbance term is given by u_{t+h}. Here, $x_{0,t}$ denotes the last available observation of the high-frequency predictor in period t, while $x_{P-1,t}$ is the lag of order $P-1$ within this period. For example, if daily predictors are used to forecast a monthly observed dependent variable, then $x_{0,t}$ corresponds to the last day in the current month t, while $x_{P-1,t}$ is the daily observation of the predictor $P-1$ days before, for $P-1=29$, say. The traditional approach replaces the potentially large number of intraperiod regressors by a single regressor, the sample average, and estimates the model by ordinary least squares (OLS). Recently, the so called Mi(xed) Da(ta) S(ampling) regression models have been suggested to explicitly allow for different frequencies at which the dependent variable and the predictor are observed. In the above MIDAS regression, a large number of parameters has to be estimated such that OLS estimates may be imprecise, and, as a consequence of the poorly estimated high-frequency weights β_p, the model may produce inaccurate forecasts. It is therefore desireable to reduce the dimensionality of the estimation problem.

To this end, Ghysels et al. (2007), among others, approximate the high-frequency weights by a parsimoniously parametrized function, such that the forecasting model becomes

$$\begin{aligned} y_{t+h} &= \alpha_0 + \alpha_1 \sum_{p=0}^{P-1} \omega_p(\theta)\, x_{p,t} + u_{t+h}, \\ \omega_p(\theta) &= \frac{\exp\left(\theta_1 \cdot p + \ldots + \theta_K \cdot p^K\right)}{\sum_{j=0}^{P-1} \exp\left(\theta_1 \cdot j + \ldots + \theta_K \cdot j^K\right)}, \end{aligned} \tag{1.1}$$

where $\theta = (\theta_1, \ldots, \theta_K)'$ is unknown. Clearly, $\omega_j(\theta) \in [0,1]$ and $\sum_{j=0}^{p} \omega_j(\theta) = 1$. The intraperiod weights thus follow a K parameter exponential Almon lag. The parameters α_0, α_1, and $\theta_1, \ldots, \theta_K$ are estimated by non-linear least squares (NLS). A parsimonious specification results for $K = 2$. Alternatively, the Beta lag distribution has been introduced by Ghysels et al. (2007).

This estimation procedure has been applied to forecast macroeconomic and financial time series to examine the potential improvement in forecast accuracy of mixed frequency models. For instance, Clements and Galvão (2009) use several monthly leading indicators to forecast quarterly U.S. output growth, showing that the MIDAS approach yields a sizeable reduction in terms of root mean squared forecast error relative to the autoregressive benchmark. Andreou et al. (2013) use daily financial variables to forecast output growth, again highlighting the advantage of MI-

DAS regressions relative to the random walk or autoregressive model.[1] In finance, Ghysels et al. (2006) predict future volatility of the Dow Jones index and thereby point out that forecasts generated by MIDAS regressions are more precise than forecasts by the autoregressive fractionally integrated benchmark model.[2] More recently, Engle et al. (2013) and Asgharian et al. (2013) extend the MIDAS approach to forecast volatility over short and long horizons by incorporating macroeconomic fundamentals as predictors of return variability.

Although the parametric approach described above may account for plausible shapes of the lag distribution, in a given estimation problem, the class of functions may not be able to represent the actual weighting function. In this chapter, the high-frequency lag distribution is therefore estimated without imposing a particular functional form. In this sense, the approach is non-parametric. The estimation procedure rests on the assumption that the unobserved weighting function is smooth in the sense that the second differences

$$\Delta^2 \beta_p = \beta_p - 2\beta_{p-1} + \beta_{p-2},\ p = 2, \ldots, P-1, \tag{1.2}$$

are small. If the second differences approximately measure the curvature of the weighting function, then the estimator suggested in this chapter penalizes large curvature, giving rise to a smooth estimate of the high-frequency weights. The estimator is conveniently obtained as a modified OLS estimator. In a companion paper, see Breitung et al. (2013), this estimator is used to forecast monthly inflation rates using daily observations of an energy price index. In this chapter, the non-parametric MIDAS regression is employed to forecast stock market volatility of the German stock index. Future return variability is sampled at a biweekly or monthly frequency and current and lagged daily absolute returns of the index are considered as predictors of this medium to long-term volatility measure. In addition to the MIDAS-PLS and parametric MIDAS-NLS approach, the GARCH(1,1) model is included in the analysis as a benchmark. Since the GARCH model uses daily observations exclusively, the in-sample fit and the forecast accuracy of the non-parametric MIDAS approach can be compared to the alternative parametric MIDAS specification as well as more traditional models that do not employ a mixed frequency scheme. Using the mean squared forecasting error (MSFE) and the mean absolute forecasting error (MAFE) as measures of forecast precision, the MIDAS-PLS approach delivers

[1] Some of the approaches employed in these papers obtain early forecasts of low-frequency variables within a given time period by incorporating all currently available information, possibly measured at different frequencies, which is also known as nowcasting. See for example Giannone, Reichlin, and Small (2008).

[2] See also Ghysels and Valkanov (2012) for a review and extensions of the MIDAS approach to forecast volatility.

more accurate forecasts than both the parametric MIDAS regresssion and the GARCH(1,1) volatility model. This observation holds in particular for forecasts of monthly volatility.

This chapter proceeds as follows. In section 1.2, the MIDAS regression is reviewed and the penalized least squares estimator is presented. The estimator depends on an a priori unspecified smoothing parameter and methods to choose this parameter are suggested. In section 1.3, the performance of the non-parametric approach relative to the parametric MIDAS estimator is studied by a Monte Carlo simulation for several shapes of weight functions that have been discussed in the literature. The goal is to examine the mean squared error (MSE) of the non-parametric and the parametric approach, stressing the relative advantages and shortcomings of the non-parametric approach to the parametric procedure. Section 1.4 uses the MIDAS-PLS estimator to forecast volatility of the German stock index DAX and compares the in-sample fit and the out-of-sample forecast performance relative to the parametric MIDAS and the GARCH(1,1) model. Section 1.5 concludes this chapter.

1.2 Penalized least squares in MIDAS regressions

1.2.1 Estimator

In this section, a non-parametric approach for estimating weight functions in MIDAS regression models is motivated within a distributed lag model as described in, among others, Shiller (1973). For simplicity, consider a model with a single high-frequency predictor for y_{t+h} and no constant,

$$y_{t+h} = \sum_{p=0}^{P-1} \beta_p x_{p,t} + u_{t+h}, \tag{1.3}$$

where the total number of high-frequency regressors, P, incorporating current and lagged values of the predictor, is given. Below, we extend this model to an empirically more relevant setup. The weights assigned to intraperiod observations are not assumed to be a member of a particular class of parametric models. However, it is required that these weights do not change abruptly in a given period of time, such that the weights can be roughly described by a curve. To incorporate this view on the parameters in (1.3), the penalized least squares objective function is studied

$$S(\lambda) = \sum_{t=1}^{T} u_{t+h}^2 + \lambda \sum_{p=2}^{P-1} \left(\Delta^2 \beta_p\right)^2, \tag{1.4}$$

where $\lambda \geq 0$ is a parameter to be chosen by the researcher and Δ^2 denotes the second difference

operator as in (1.2). In this section, the parameter λ is taken as given. In section 1.2.2 methods to select λ are discussed.

This objective function trades off a goodness-of-fit criterion with a smoothness requirement for the high-frequency weights, where smoothness is measured in terms of second differences of the parameters. For $\lambda = 0$, we obtain the OLS estimator, which imposes no smoothness on the weighting function. In the limit as $\lambda \to \infty$, the penalty term dominates the sums of squared residuals in the objective function, forcing the second difference of the weights to equal the same constant. Hence, the weighting function becomes linear. For intermediate values $0 < \lambda < \infty$, the weighting function attains a smooth shape between these two extremes.

To rewrite this problem more concisely, let the $(P-2) \times P$ matrix D by given by

$$D = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 & -2 & 1 \end{bmatrix}. \tag{1.5}$$

Let $y = (y_{1+h}, \ldots, y_{T+h})'$, and $X = [x_1, \ldots, x_T]'$, where $x_t = (x_{0,t}, \ldots, x_{P-1,t})'$. Then the minimization problem becomes

$$\min_{\beta \in \mathbb{R}^P} S(\lambda) = \min_{\beta \in \mathbb{R}^P} (y - X\beta)' (y - X\beta) + \lambda \, \beta' D' D \beta, \tag{1.6}$$

with $\beta = (\beta_0, \ldots, \beta_{P-1})'$. The solution to (1.6) is given by

$$\widehat{\beta}_\lambda = (X'X + \lambda \, D'D)^{-1} X'y = \left((X'X/T) + \overline{\lambda} \, D'D\right)^{-1} (X'y/T), \tag{1.7}$$

with $\overline{\lambda} = \lambda/T$, provided that $(X'X + \lambda D'D)$ is non-singular. We refer to estimator as the penalized least squares (PLS) estimator, or the MIDAS-PLS regression, in contrast to the parametric MIDAS-NLS approach. Clearly, for $\lambda = 0$, the estimator reduces to the OLS estimator. Moreover, by an algebraic rule for matrix inversion (see for example Lütkepohl (1996), page 29),

$$(X'X + \lambda D'D)^{-1} = (X'X)^{-1} - (X'X)^{-1} D' \left(D (X'X)^{-1} D' + \frac{1}{\lambda} I_{P-2} \right)^{-1} D (X'X)^{-1}. \tag{1.8}$$

Thus, as $\lambda \to \infty$, $\widehat{\beta}_\lambda$ approaches the restricted least squares estimator subject to $D\beta = 0$. These

restrictions in turn imply that the weights lie on a straight line.[3] Several additional comments about this estimator can be made.

First, each component of $\widehat{\beta}_\lambda$ is a weighted sum of the OLS coefficients,

$$\widehat{\beta}_\lambda = \left(I_P + \lambda \left(X'X\right)^{-1} D'D\right)^{-1} \widehat{\beta}, \tag{1.9}$$

and a larger value of λ imposes more smoothness on the weights. Hence the penalized estimator can be viewed as a smoothed OLS estimator and λ as a smoothing parameter.

Second, following Shiller (1973), the estimator is conveniently obtained from the regression

$$y^* = X^* \widehat{\beta}_\lambda + \widehat{u}^*, \tag{1.10}$$

with the $(T + (P-2)) \times 1$ vector $y^* = \left(y', 0'_{P-2}\right)'$ and the $(T + (P-2)) \times P$ matrix $X^* = \left[X', \sqrt{\lambda} D'\right]'$. Thus, the estimator is easy to compute once the smoothing parameter has been chosen.

Third, as noted by Anup and Maddala (1984) for example, since the smoothed OLS estimator arises from the penalized least squares objective function (1.6), it can be viewed as a generalized ridge estimator in which the matrix $D'D$ replaces the identity matrix of the ordinary ridge estimator (see Hoerl and Kennard (1970)). In this view, the penalized least squares estimator is a generalized shrinkage estimator, and λ determines the degree of shrinkage. The shrinkage terminology stems from the fact that the penalized least squares estimator considered in this chapter is equivalently characterized as

$$\begin{aligned} \widehat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^P}{\operatorname{argmin}} & \left(y - X\beta\right)' \left(y - X\beta\right) \\ \text{s.t. } & \beta' D' D \beta \leq d, \end{aligned} \tag{1.11}$$

where there is a one-to-one correspondence between λ and $d \geq 0$.[4] This interpretation may be useful in MIDAS regressions. Ridge regression is able to reduce the expected squared loss relative to OLS in particular when the regressor matrix X is nearly collinear (see, for example, Hoerl and Kennard (1970)). In empirical applications, non-negligible correlation between the regressors may be encountered in MIDAS regressions, if, for instance, daily or weekly observed predictors are used. In this case, ridge regression provides more precise, albeit biased, estimates

[3] See Chipman (2011), page 314-315, for a derivation.

[4] Problems (1.6) and (1.11) are equivalent in that for each $0 \leq \lambda < \infty$, there exists a $d \geq 0$ such that the problems have the same solution, and vice versa, see Fu (1998).

of the weighting function than OLS. Although this reasoning may be used for the ordinary ridge regression, the Monte Carlo evidence for the generalized ridge estimator considered here agrees with this notion, see section 1.3.1.

Finally, an extension of the simple model (1.3) to include a constant and lagged variables of the low-frequency process is easily obtained. In this way, the simple MIDAS regression from above is extended to the ADL-MIDAS model employed by Andreou et al. (2013). Let $L-1$ denote the number of lags of the dependent variable. Then the MIDAS regression model becomes

$$y_{t+h} = \mu + \sum_{l=0}^{L-1} \alpha_l y_{t-l} + \sum_{p=0}^{P-1} \beta_j x_{p,t} + u_{t+h}.$$

In matrix notation

$$y = X_1\alpha + X_2\beta + u = X\gamma + u,$$

where the $T \times (L+1)$ matrix X_1 has rows $(1, y_t, \ldots, y_{t-L+1})$, $\alpha = (\mu, \alpha_0, \ldots, \alpha_{L-1})'$, $X = [X_1,\ X_2]$ and $\gamma = (\alpha', \beta')'$. Consider the problem

$$\min_{\gamma \in \mathbb{R}^q} S(\lambda) = (y - X\gamma)'(y - X\gamma) + \lambda\, \beta' D' D \beta, \tag{1.12}$$

with $q = P + L + 1$. Notice that only the coefficients of the high-frequency predictors are penalized. Let $M = I_T - X_1 (X_1'X_1)^{-1} X_1$ and $\widetilde{X}_2 = MX_2$. Then, using the results in Farebrother (1978), the solution to the above problem is

$$\widehat{\gamma}_\lambda = \begin{bmatrix} \widehat{\alpha}_\lambda \\ \widehat{\beta}_\lambda \end{bmatrix} = \begin{bmatrix} (X_1'X_1)^{-1} X_1' \left(y - X_2\widehat{\beta}_\lambda\right) \\ \left(\widetilde{X}_2'\widetilde{X}_2 + \lambda D'D\right)^{-1} \widetilde{X}_2' y \end{bmatrix}.$$

Hence the weighting function $\widehat{\beta}_\lambda$ is obtained as a shrinkage estimator once the low-frequency variables have been partialled out. Notice that the estimate of α is affected by the degree of smoothing via $\widehat{\beta}_\lambda$. In MIDAS regressions, it is natural to restrict smoothing to the high-frequency predictors and to estimate the low- frequency parameters simply by OLS. This modification is achieved by replacing X with $\widetilde{X} = \left[X_1,\ \widetilde{X}_2\right]$.

1.2.2 Choice of the smoothing parameter

In this section, several selection procedures for the smoothing parameter λ are reviewed. These include a plug-in estimator, information criteria and a simulation-based approach.

Plug-in estimator

Following Anup and Maddala (1984), in this section we study a departure from the fixed parameter model we have considered so far to motivate a plug-in estimator for the smoothing parameter. Suppose that the second differences satisfy

$$\begin{aligned} D\beta &= v, \\ v|X &\sim \left(0, \sigma_v^2 I_{P-2}\right), \end{aligned}$$

where the regression error u and v are uncorrelated. The above framework can be viewed as an example of imposing stochastic restrictions on β, such that the second differences are random quantities.[5] Now consider the artificial regression

$$y^* = X^*\beta + \epsilon^*,$$

with $y^* = \left(y', 0'_{P-2}\right)'$, $X^* = [X', -D']'$, and $\epsilon^* = (u', v')'$. Notice that

$$\Omega = \mathbb{E}\left[\epsilon^* \epsilon^{*\prime} \middle| X\right] = \begin{bmatrix} \sigma_u^2 I_T & 0 \\ 0 & \sigma_v^2 I_{P-2} \end{bmatrix} = \sigma_u^2 \begin{bmatrix} I_T & 0 \\ 0 & \lambda^{-1} I_{P-2} \end{bmatrix},$$

with $\lambda = \sigma_u^2/\sigma_v^2$, and suppose that this ratio is known. Then the generalized least squares estimator of β coincides with the PLS estimator. To estimate λ in this model, let $\widehat{\lambda} = \widehat{\sigma}_u^2/\widehat{\sigma}_v^2$, with $\widehat{\sigma}_u^2 = \widehat{u}'\widehat{u}/(T-P)$ where $\widehat{u} = \left(I - X\left(X'X\right)^{-1}X'\right)y$. Regarding σ_v^2, a natural choice is $\widehat{\beta}' D' D \widehat{\beta}/(P-2)$.

Information criterion

The fitted values resulting from the smoothed OLS estimation are

$$\widehat{y}_\lambda = X\left(X'X + \lambda D'D\right)^{-1} X'y = Z_\lambda y,$$

with $Z_\lambda \equiv X\left(X'X + \lambda D'D\right)^{-1} X'$. Notice that for $\lambda = 0$, Z_λ reduces to the familiar orthogonal projection matrix in a linear regression model. Consquently, its trace is equal to P. In general, define $\kappa_\lambda \equiv \text{tr}\left[Z_\lambda\right]$. It can be shown that κ_λ is monotonically decreasing in λ and that κ_λ approaches two as $\lambda \to \infty$.[6] The trace indicates the pseudo dimension of the parameter vec-

[5] In the sense of Goldberger and Theil (1961), additional information is available about the parameters that is stochastic in nature.

[6] This limit is easily established using (1.8).

tor: for $\lambda = 0$, the number of parameters is equal to P. For $\lambda \to \infty$, the number of parameters reduces to two. If $\lambda > 0$, the non-parametric estimator achieves a dimension reduction relative to OLS. Since non-integer values can occur, the term pseudo dimension is used.

The trade-off between goodness-of-fit (in terms of squared loss) and dimension reduction of the non-parametric estimation can be exploited to construct an information criterion to choose the smoothing parameter. Following Hurvich, Simonoff, and Tsai (1998), the modified Akaike criterion is

$$\text{AIC}(\lambda) = \log\left[\text{SSR}(\lambda)\right] + \frac{2(\kappa_\lambda + 1)}{T - \kappa_\lambda - 2},$$

where $\text{SSR}(\lambda) = (y_h - \widehat{y}_\lambda)'(y_h - \widehat{y}_\lambda)$. The degree of smoothing is selected by minimizing this criterion with respect to λ. This task can be attempted by grid search. Alternatively, as shown in appendix 1.A, by employing brute force minimization, the minimizer $\lambda^\star_{\text{AIC}}$ results as the solution to the first-order condition

$$\begin{aligned}&\left(y_h' X Q^{-1}_{\lambda^\star_{\text{AIC}}} (D'D) Q_{\lambda^\star_{\text{AIC}}} X' \left(\left(I - Z_{\lambda^\star_{\text{AIC}}}\right) y_h\right)\right) \left(\text{SSR}(\lambda^\star_{\text{AIC}})\right)^{-1} \\ &= (T-1)\,\text{tr}\left[(X'X) Q^{-1}_{\lambda^\star_{\text{AIC}}} (D'D) Q^{-1}_{\lambda^\star_{\text{AIC}}}\right] \left(T - \kappa_{\lambda^\star_{\text{AIC}}} - 2\right)^{-2},\end{aligned}$$

where $Q_\lambda = (X'X + \lambda D'D)$. This condition can be solved numerically.

Cross validation

Adapting the discussion in Golub et al. (1979), cross validation can be used to estimate the smoothing parameter and the procedure is summarized briefly. Here, the sample is partitioned into B blocks. Then a value of the smoothing parameter λ is fixed. Using the observations in $B-1$ blocks, $\widehat{\beta}_\lambda$ is computed, where the b-th block is left out for an out of sample evaluation, for $b = 1, 2, \ldots, B$. Let

$$\epsilon_b(\lambda, T_b) = \frac{1}{T_b} \sum_{t=1}^{T_b} \left(y^{(b)}_{t+h} - \widehat{\beta}'_\lambda x^{(b)}_t\right)^2,$$

where T_b is the number of observations in block b, $y^{(b)}_{t+h}$ is observation t in block b and analogous for $x^{(b)}_t$. The cross validation (CV) criterion is

$$\text{CV}(\lambda) = \frac{1}{B} \sum_{b=1}^{B} \epsilon_b(\lambda, T_b).$$

Repeating this procedure for a range of values of the smoothing parameter gives the CV objective function as a function of these values. The minimizer of this objective function is the optimal value according to cross validation. Common choices for the number of blocks are $B = 5$ or $B = 10$.

These methods can provide a guideline to selecting the smoothing parameter. The rule-of-thumb estimator is simple to compute but somewhat ad hoc. The CV method can be computationally demanding if the range of possible candidates to evaluate is large. The AIC criterion is therefore selected henceforth. It should be noted however that selecting λ involves some leeway and the robustness of the results to different choices of λ should be considered in practice.

1.3 Monte Carlo experiments

In this section, we compare the small-sample properties of the non-parametric and parametric approach to estimating MIDAS regression models.

1.3.1 Simulation design and evaluation criteria

Consider the data-generating process

$$y_{t+h} = \sum_{p=0}^{P-1} \beta_p x_{p,t} + u_{t+h}, \tag{1.13}$$

$$\beta_p = \alpha_1 \, \omega_p(\theta), \tag{1.14}$$

where u_{t+h} is independently standard normally distributed. Several alternative specifications for the true shape of the weights β_p in (1.14) are examined in the following sections and are presented separately.

The high-frequency predictor is generated by the AR(1) process

$$\begin{aligned} x_{p,t} &= \psi x_{p-1,t} + \left(1 - \psi^2\right)^{1/2} \epsilon_{p,t}, \\ \epsilon_{p,t} &\overset{iid}{\sim} \mathcal{N}\left(0, 1\right), \end{aligned}$$

for $p = 0, \ldots, P-1$ and $\psi \in \{0.2, 0.4, 0.8\}$. The sample size is $T = 100$ and the number of high-frequency regressors is $P = 30$. This number of intraperiod regressors is taken as given, assuming that an appropriate choice regarding the lag length in the forecasting model has been made. Moreover, the uncentered R^2 is either 0.3 or 0.15, which intends to illustrate

the performance of both methods in empirically relevant settings, see for example Breitung et al. (2013). To do so, the scaling parameter α_1 is selected by grid search such that for each specification the average R^2 across simulations is fixed at the desired level.

The MIDAS regression (1.14) is estimated by MIDAS-PLS and by MIDAS-NLS with exponential Almon lag polynomial of order $K = 2$, which is a popular choice in applications of the parametric MIDAS approach, see for example Clements and Galvão (2009) or Monteforte and Moretti (2013). The number of Monte Carlo replications is 200, which is rather small, but due to the fact that both the non-linear least squares optimization of MIDAS-NLS and determining the smoothing parameter by AIC repeatedly for MIDAS-PLS can be time consuming. The simulation results are thus taken as first evidence of the performance of the two procedures and can be extended in future research.

The estimation methods are compared in terms of their in-sample fit and their out-of-sample forecasting performance. First, the in-sample fit is is measured by the unconditional MSE, that is the median MSE across Monte Carlo simulations of MIDAS-PLS relative to the median MSE of parametric MIDAS. Hence a MSE ratio below one implies that MIDAS-PLS has smaller MSE than parametric MIDAS. The modified AIC criterion is used for selecting the smoothing parameter for MIDAS-PLS. In addition, as the true weights in (1.14) are known, simulation results are reported for the choice of the smoothing parameter that minimizes expected mean squared error conditional on X. From (1.9),

$$\mathbb{E}\left[\widehat{\beta}_\lambda \middle| X\right] - \beta = \left(\left(I_P + \lambda \left(X'X\right)^{-1} D'D\right)^{-1} - I_P\right)\beta = \Psi\left(\lambda\right)\beta,$$

with $\Psi\left(\lambda\right) = \left(\left(I_P + \lambda \left(X'X\right)^{-1} D'D\right)^{-1} - I_P\right)$, and since X is independent of u with $\mathbb{E}\left[uu'\right] = \sigma_u^2 I_T$,

$$Var\left[\widehat{\beta}_\lambda \middle| X\right] = \sigma_u^2 \Psi\left(\lambda\right)\left(X'X\right)^{-1}\Psi\left(\lambda\right)'.$$

The MSE optimal choice for λ conditional on X is then obtained by minimizing

$$\text{tr}\left[\Psi\left(\lambda\right)\beta\beta'\Psi'\left(\lambda\right)\right] + \sigma^2 \text{tr}\left[\Psi\left(\lambda\right)\left(X'X\right)^{-1}\Psi\left(\lambda\right)'\right].$$

The minimizer is obtained numerially and is denoted λ_{MSE}. In the simulation experiment the shrinkage parameter is reported relative to sample size, that is, $\overline{\lambda}_{\text{MSE}} = \lambda_{MSE}/T$. Similarly,

$\overline{\lambda}_{\text{AIC}} = \lambda_{\text{AIC}}/T$ is computed by minimizing the modified AIC criterion. In addition, a grid of values of $\overline{\lambda}$ is provided that allows to assess the sensitivity with respect to the choice of the smoothing parameter.

Second, the accuracy of out-of-sample forecasts made by the non-parametric and the parametric approach in this simple model is evaluated. To this end, we partition the whole sample into an estimation sample, comprising observations $1, 2, \ldots, T^e$, and a forecasting sample including observations indexed by $T^e + 1, \ldots, T$. Let T^f denote the sample size of the forecasting sample such that $T = T^e + T^f$. The estimation sample is used to obtain a baseline estimate of the weights β_p,

$$\widehat{y}_{t+h} = \sum_{p=0}^{P-1} \widehat{\beta}_p x_{t,p} + \widehat{u}_{t+h},$$

for $t = 1, \ldots, T^e$,where $\widehat{\beta}_p$ is a suitable estimator of the weights. The one-step ahead forecast is

$$y^f_{T^e+1+h} = \sum_{p=0}^{P-1} \widehat{\beta}_p x_{T^e+1,p}.$$

The estimation sample is then enlarged by one observation and the model is reestimated using observations $1, 2, \ldots, T^e + 1$ to produce the next one-step ahead forecast. Proceeding in this fashion, a series of one-step ahead forecast errors is obtained. The resulting mean squared forecasting error (MSFE) is

$$\frac{1}{T^f} \sum_{t=1}^{T^f} \left(y_{T^e+h+t} - y^f_{T^e+h+t} \right)^2.$$

This forecasting exercise is replicated across simulations. The forecasting accuracy is compared via the median mean squared forecasting errors of MIDAS-PLS relative to MIDAS-NLS. Again, a ratio below one indicates that non-parametric MIDAS yields more precise forecasts on average.

1.3.2 Exponentially and linearly declining weights

The first specification fits into the parametric MIDAS framework. The weights are exponentially declining

$$\omega_p = \frac{\exp(\theta_1 p)}{\sum_{j=0}^{P-1} \exp(\theta_1 j)}, \quad p = 0, \ldots, P-1. \tag{1.15}$$

Figure 1.2 (see appendix 1.B) shows the true parameter function when $\theta_1 = -0.1$ (slow decay) and $\theta_1 = -0.2$ (moderate decay). Table 1.3 in appendix 1.B shows the ratio of MSEs of the non-parametric estimator relative to the two-parameter exponential Almon specification. With exponentially declining weights, the relative performance of the non-parametric estimator hinges on two features: the shape of the weighting function and the signal-to-noise ratio as measured by R^2. For more rapid exponential decay and $R^2 = 0.3$, the parametric approach performs better than the non-parametric estimator. For slower exponential decay, however, the non-parametric estimator outperforms the parametric MIDAS approach. For small values of R^2, the PLS estimator estimates the weights more precisely, in particular if the autocorrelation of the regressor is large, implying high collinearity. Notice that the AIC criterion suggests choices that are close to the MSE optimal value for the smoothing parameter. These results illustrate the potential of this estimation method if the true weighting function declines moderately.

As another example of a monotonically declining weight function, consider the linear function

$$\omega_p = \frac{a_0 + a_1 p}{a_0 P + a_1 (P-1) P/2}. \tag{1.16}$$

The weights are normalized to sum up to unity. We consider two examples: a linear declining weight function and a near-flat weight function, see figure 1.3 in appendix 1.B. Table 1.4 (see appendix 1.B) shows ratios of MSEs in this case. It turns out that the non-parametric estimator clearly outperforms the parametric MIDAS approach with $K = 2$. Note that the PLS estimator approaches a straight line as $\lambda \to \infty$. Accordingly, the nonparametric approach has an advantage relative to the standard MIDAS approach as the exponential Almon lag is misspecified in this case.

The ratios of the out-of-sample forecast errors (MSFE) are always close to unity. This suggest that both approaches perform similarly in this one-step ahead forecasting competition, albeit the non-parametric approach tends to provide slightly more reliable forecasts.

1.3.3 Hump-shaped weights

Next, we examine the hump-shaped pattern

$$\omega_p = \frac{\exp\left(\theta_1 p - \theta_2 p^2\right)}{\sum_{j=0}^{P-1} \exp\left(\theta_1 j - \theta_2 j^2\right)}. \tag{1.17}$$

The weight function is depicted in figure 1.4 (see appendix 1.B) for different choices of θ_1 and θ_2. In both cases, the weight function attains a maximum at lag 5. The two specifications differ in the degree of curvature of the weight function before and after the peak. Again, the scaling parameter α_1 is selected corresponding to $R^2 = 0.3$ and $R^2 = 0.15$. Table 1.5 in appendix 1.B shows the MSE ratios in this case.

The non-parametric estimator works well for moderate curvature. If the hump shape is more pronounced, the parametric approach provides more accurate estimates according to the MSE ratios. The relative performance of the non-parametric estimator depends on the degree of smoothing, but the AIC produces useful, although not optimal, choices. Furthermore, the out-of-sample MSFE of MIDAS-PLS is slightly lower than one.

1.3.4 Cyclical weights

To point out the flexibility of the non-parametric approach, consider a cyclical weight function that is difficult to tackle within the parametric MIDAS framework,

$$\omega_p = \frac{1}{P}\left[1 + \sin\left(\frac{2\pi p}{P-1}\right)\right]. \tag{1.18}$$

Figure 1.5 in appendix 1.B shows the shape of the weighting function. Table 1.6 displays the relative performance of the non-parametric estimator in this case, see appendix 1.B. In general, the non-parametric approach estimates the weight function more accurately, and this observation holds in particular for $R^2 = 0.15$. Due to the unusual form of the weights, it is examined whether the three-parameter exponential Almon lag improves upon the two-parameter specification. However, the three parameter specification does not yield more accurate estimates. This example hence illustrates the potential benefits of leaving the functional form unrestricted. As before, the out-of-sample MSFE are marginally smaller than unity, suggesting a slightly better performance of the non-parametric approach.

1.4 Forecasting volatility with MIDAS regressions

In this section, the penalized MIDAS regression is employed to forecast volatility of the German stock market. Return volatility is a key variable for portfolio allocation and risk management. For instance, Campbell and Viceira (2001) derive the optimal portfolio weight for an investor with constant relative risk aversion that chooses between a riskless asset and stocks, showing that the optimal weight is function of predicted future return volatility. In practice, volatility also serves as an input variable for assessing market risk; see for example Engle (2004) for a simple illustration of how estimated future volatility is used to compute the so called value-at-risk, a widely used measure in risk management.

The volatility forecasting literature is large with theoretical contributions suggesting statistical models of volatility and empirical applications. We refer the reader to Poon and Granger (2003) for an extensive review. Here, our focus is on how the MIDAS scheme can be exploited to forecast volatility and how the estimation procedure suggested in this chapter can be used to do so.

1.4.1 Data

In this exercise, the German stock index DAX is considered. The observations are daily closing prices of the DAX stock index. The daily series starts on January 1st 1965 and ends on May 14th 2013 and was obtained from Datastream. Figures 1.6 and 1.7 in appendix 1.C plot the DAX index and the daily log returns for the entire sample period and table 1.7 presents sample statistics for the full sample and the subsamples beginning in January 1993 and 2002. These statistics illustrate some of the typical characteristics of stock return series indicating that the return distribution is mostly negatively skewed and fat-tailed. One noticeable feature the data is the larger variability of the series since the mid 1980s.

To illustrate the tools presented in previous sections, attention is restricted to the subsample starting in January 1993, covering the most recent twenty years of daily return data. In addition to the post-reunfication sample, the return series starting in 2002, which considers the period after the stock market bubble in 2001 and the events of 9/11, both which affected the German economy, is studied as well.

1.4.2 Sampling scheme and notation

To explain the MIDAS structure in this application, the sampling scheme and measures of return variability are discussed, following the notation in Alper et al. (2012) and Ghysels et al. (2006) for ease of comparison. It is important to distinguish the low-frequency process to be forecasted and the high-frequency predictors. To this end, let time index t denote biweekly or monthly sampling, respectively. Here, the compounded return from time $t-1$ to time t is denoted as $r_{t,t-1}$. Within this given time period, there are m equidistant daily returns available, where $m = 10$ and $m = 20$ for biweekly and monthly sampling, respectively.

The return at the end of time t is

$$r_{t,t-1} = \log\left(P_t^{(m)}\right) - \log\left(P_{t-1}^{(m)}\right), \tag{1.19}$$

where $P_t^{(m)}$ is the daily closing price of the stock at the end of period t. To define the dependent variable and the predictors in the MIDAS regressions specified below, it is convenient to introduce daily returns explicitly as

$$r^{(m)}_{t-(j-1)/m,t-j/m} = \log\left(P^{(m)}_{t-(j-1)/m}\right) - \log\left(P^{(m)}_{t-j/m}\right),\ j = 1,\ldots,m, \tag{1.20}$$

where $P^{(m)}_{t-j/m}$ is the closing price of the stock on day $m-j$ in period t. Clearly, $r_{t,t-1}$ is the sum of the daily returns

$$r_{t,t-1} = \sum_{j=1}^{m} r^{(m)}_{t-(j-1)/m,t-j/m}, \tag{1.21}$$

consistent with (1.19) and (1.20).

We aim to predict the variability of future returns for some forecasting horizon h. Following Brooks and Persand (2003) and Alper et al. (2012) the measure of future variability of the DAX returns employed in this study is given by

$$RV_{t+h,t} = \sum_{j=1}^{hm} \left[r^{(m)}_{t+h-(j-1)/m,t+h-j/m}\right]^2, \tag{1.22}$$

which is also known as realized variance (at horizon h).[7] We consider $RV_{t+h,t}$ as the dependent

[7] Conditional on information available up to time $t-1$, $\mathbb{E}_{t-1}\left[r_{t-(j-1)/m,t-j/m}\right] \approx 0$, where $\mathbb{E}_{t-1}[\cdot]$ denotes the conditional expectation given the information set at $t-1$, and that returns are conditionally uncorrelated. In this case $Var_{t-1}[r_{t,t-1}] = Var_{t-1}\left[\sum_{j=1}^{m} r_{t-(j-1)/m,t-j/m}\right] \approx \sum_{j=1}^{m} \mathbb{E}_{t-1}\left[r_{t-(j-1)/m,t-j/m}\right]^2$, see Hansen and Lunde (2005), page 878, for example.

variables in the MIDAS regressions below.[8] For instance, with monthly sampling, $r_{t,t-1}$ denotes the monthly return which can be decomposed into $m = 20$ daily returns as in (1.21). Accordingly, with $h = 1$, say, $RV_{t+1,t}$ is computed using non-overlapping daily returns as in (1.22) and serves as a proxy for return varibility in the following month. Andersen and Bollerslev (1998) or Andersen et al. (2001) argue that although other unbiased estimators of the conditional return variance are available, realized variance provides a more accurate approximation of the unobserved return variability. We then run a MIDAS regression using daily predictors, which are available up to month t, to produce forecasts of this measure of return variability.

1.4.3 MIDAS volatility regressions

We adapt the notation of Ghysels et al. (2006) to formulate the MIDAS regression

$$RV_{t+h,t} = \mu + \sum_{p=0}^{P-1} \beta_p x^{(m)}_{t-p/m,t-(p+1)/m} + u_{t+h,t}, \tag{1.23}$$

where $RV_{t+h,t}$ is defined in (1.22) and $x^{(m)}_{t-p/m,t-(p+1)/m}$ is a potential high-frequency predictor. In the MIDAS volatility regression, for $p = 0$, $x^{(m)}_{t,t-1/m}$ is the latest available observation of the daily predictor, while $x^{(m)}_{t-(P-1)/m,t-P/m}$ denotes the daily predictor at lag $P - 1$. Hence low-frequency future volatility, measured biweekly or monthly, say, is forecasted using P current and lagged observations of the high-frequency predictors. Due to the different sampling of the dependent variable and the predictor, the MIDAS scheme arises.

We investigate one daily predictor, lagged absolute returns, which have been suggested as measures of past volatility.[9] Alternatively, squared returns can be employed as predictors, but building on the evidence in Ghysels et al. (2006), absolute returns perform better in a mean squared error sense. Hence the MIDAS regressions is

$$RV_{t+h,t} = \mu + \sum_{p=0}^{P-1} \beta_p \left| r^{(m)}_{t-p/m,t-(p+1)/m} \right| + u_{t+h,t}. \tag{1.24}$$

The daily weights β_p can be estimated using the parametric or the non-parametric approach. When employing the PLS estimator, the smoothing parameter is chosen by the AIC criterion. For the parametric specification, we follow Ghysels et al. (2006) and Alper et al. (2012) and use

[8] Other authors stress that realized variance depends on m, as in $RV^{(m)}_{t+h,t}$, for example. We use the short-hand notation $RV_{t+h,t}$.

[9] See Forsberg and Ghysels (2007) for a theoretical motivation for using absolute returns as predictors of future volatility.

the two-parameter Beta lag structure

$$\begin{aligned} \beta_p &= \alpha_1 \omega_p, \\ \omega_p &= \frac{f\left(\frac{p}{P-1}; \theta_1, \theta_2\right)}{\sum_{j=1}^{P-1} f\left(\frac{j}{P-1}; \theta_1, \theta_2\right)}, \\ f(z, a, b) &= \frac{z^{a-1}(1-z)^{b-1}}{F(a, b)}, \end{aligned}$$

where $F(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $\Gamma(\cdot)$ denotes the Gamma function. This lag structure allows for various lag distributions, including moderate or fast decay when $\theta_1 = 1$ and $\theta_2 > 1$.

1.4.4 In-sample fit

The in-sample fit of the MIDAS regression model using the parametric Beta lag and the non-parametric penalized least squares approach is compared. In the MIDAS volatility regression, Ghysels et al. (2006) recommend including up to fifty lags of the predictor, and Alper et al. (2012) follow this suggestion. Here, we set $P = 50$ as the default choice. As the estimation results of these studies suggest that weights assigned to distant lags have a very small magnitude, the in-sample fit is also examined when $P = 30$ and $P = 40$. Figures 1.1a - 1.1c show the estimated lag distribtutions β_p in (1.24) for lag lengths $P \in \{30, 40, 50\}$ in the sample period beginning in January 1993. These figures illustrate the differences between the parametric and non-parametric approach in this application: the parametric estimates (straight black line) put almost all weight on the most recent 5 to 10 days. The non-parametric approach (straight blue line) follows the raw OLS estimates (dotted line) more closely. The penalized least squares approach hence results in somewhat different shapes of the lag distributions with positive weights assigned to lags as far as 25 or 50 days. For instance, with 30 lagged absolute returns, the non-parametric lag distribution peaks around days 2 to 4, declines afterwards and rises around days 16 or 17 again. Therefore, in this application the non-parametric approach produces estimates that lie between the raw, jagged OLS estimates and the relatively fast decaying NLS estimates. Whether this more flexible non-parametric approach also improves forecasting performance is studied in the next subsection.

Table 1.1 presents MSE ratios of PLS relative to parametric MIDAS for monthly and biweekly sampling, such that a ratio below one indicates a lower MSE of the PLS approach. The results are reported for realized variance $RV_{t+h,t}$ for a $h = 1$, such that next month's volatility, say, is

Table 1.1: In-sample MSE ratios

Sample		monthly sampling, $h = 1$			biweekly sampling, $h = 1$		
		$P = 50$	$P = 40$	$P = 30$	$P = 50$	$P = 40$	$P = 30$
1993 - 2013							
$(\mathbf{T = 255})$	MSE ratio	0.21	0.97	0.22	0.95	0.96	0.97
	κ_λ	13.2	7.0	8.6	12.1	11.0	8.2
2002 - 2013							
$(\mathbf{T = 142})$	MSE ratio	0.91	0.99	0.93	0.92	0.93	0.94
	κ_λ	11.5	5.2	7.6	14.0	10.4	7.6

Note: Entries are the MSE ratios of MIDAS-PLS to MIDAS-NLS based on a regression of (non-overlapping) realized variance RV_{t+h} (see (1.22)) on a constant and current and lagged absolute returns. The smoothness parameter is selected by AIC. Here, κ_λ denotes the pseudo dimension of MIDAS-PLS: if the smoothness parameter is zero, $\kappa_\lambda = P$, while $\kappa_\lambda \to 2$ as the smoothing parameter tends to infinity.

forecasted by current and lagged absolute returns. Given the differences in the estimated lag distributions presented in figures 1.1a - 1.1c, the ratios are reported for different lag lengths. The PLS approach yields a substantially better fit in some cases, with ratios varying between 0.21 and 0.93 depending on the volatility measure and the lag length. The variability in the relative performance may be explained by the observation that in some cases, the estimated lag distribution from the parametric approach differs strongly from the shape of the OLS estimates, in contrast to the non-parametric estimation method, see also figures 1.1a - 1.1c. In addition to the MSE ratios, the pseudo dimension κ_λ is reported. The parametric approach is parsimonious as it describes the lag distribution with three parameters (θ_1,θ_2 and the scaling parameter α_1, excluding the constant) and thus achieves a considerable dimension reduction relative to OLS with P parameters (excluding the constant). The non-parametric approach also reduces dimensionality relative to OLS, albeit not as strongly as the parametric procedure. Although less parsimonious, PLS provides a better fit. In the biweekly case, PLS still provides smaller in-sample MSE, although differences in the relative performance are smaller.

1.4.5 Out-of-sample forecasting exercise: design

We now turn to forecasting the volatility of the DAX index. The design of the forecast experiment follows Brooks and Persand (2003) and Raunig (2006). The samples are split into an estimation sample and a forecasting sample. The estimation sample covers the periods from January 1993 to December 2003 and the forecasting period is then January 2004 until May 2013. Forecasts are obtained recursively with a moving estimation window. Given the estimated lag

Figure 1.1a: Estimated lag distributions 1993-2013 with 50 lags

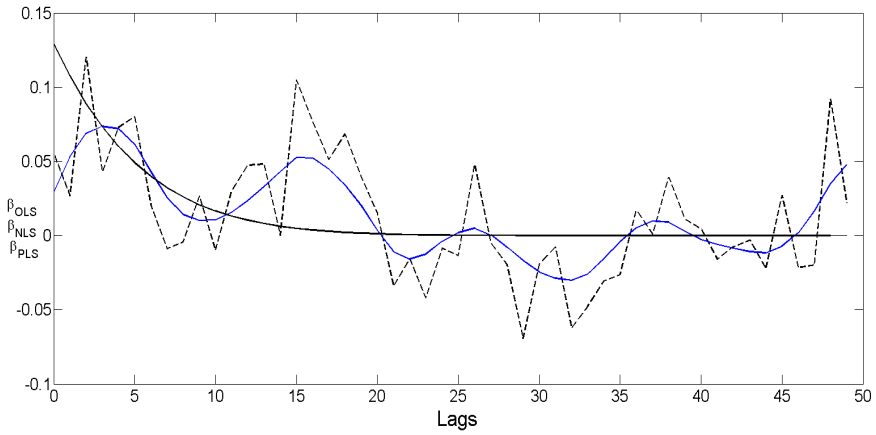


Figure 1.1b: Estimated lag distributions 1993-2013 with 40 lags

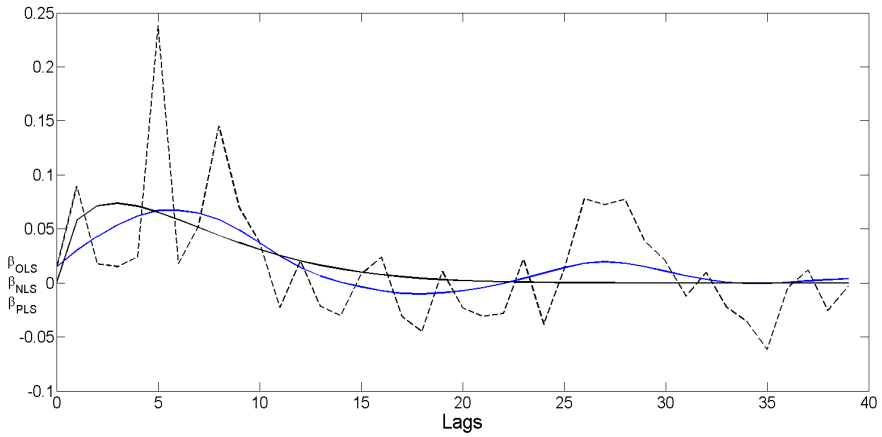
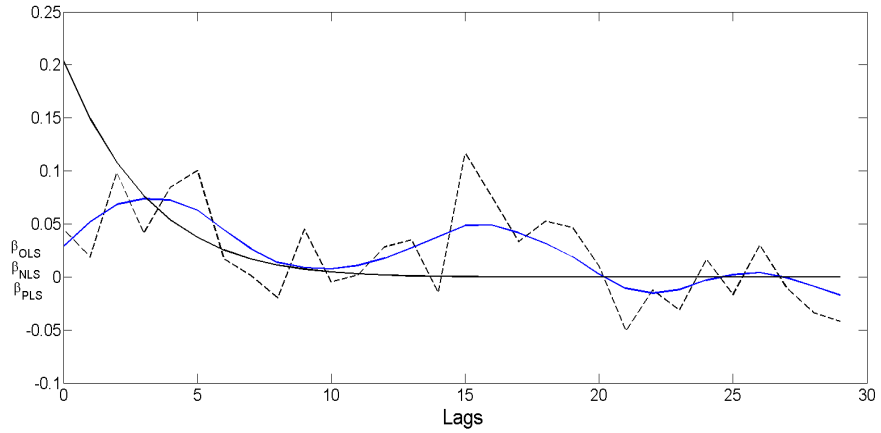


Figure 1.1c: Estimated lag distributions 1993-2013 with 30 lags



Note: OLS (dotted line), NLS (straight black line) and PLS (straight blue line) estimates for monthly sampling with $RV_{t+1,t}$ as the dependent variable and $T = 255$ observations.

distributions obtained from the estimation sample, the first one-period-ahead forecast is produced. Then the first observation from the estimation sample is dropped while the end of the sample is extended to include the observations in the next period. The MIDAS regressions are reestimated and volatility in the next period is forecasted. This forecasting scheme continues until the penultimate period in the sample is reached and volatility is forecasted for the last period in the sample.

It is important to recall that a meaningful forecast of future volatility should be non-negative. One advantage of the parametric approach using the Beta lag distribution is that the estimated lag distributions are ensured to be non-negative, which in turn leads to non-negative volatility forecasts. The penalized least squares estimates, however, are unrestricted and can be negative, leading potentially to useless volatility forecasts. To make sure that the non-parametric approach generates non-negative forecasts, recall from section 1.2 (see (1.10) and the surrounding discussion) that for a given value of the smoothing parameter, the PLS estimator can obtained by regressing $y^* = [y',\ 0'_{P-2}]'$ on $X^* = \left[X',\ \sqrt{\lambda}D'\right]'$, where in this case y denotes the vector of the volatility measure (realized volatility, say) and X is the matrix of past absolute returns, while D is the difference matrix introduced in section 1.2. We then complement this auxiliary regression by a non-negativity constraint

$$\begin{aligned} y^* &= X^*\beta_\lambda + \epsilon^* \\ &\text{s.t. } \beta_\lambda \geq 0, \end{aligned} \tag{1.25}$$

and estimate this regression to obtain the non-negative least squares estimator, see for example Liew (1976), which is readily available in MATLAB, for instance. Alternatively, the inequality constrained ridge estimator suggested by Toker et al. (2013) can be employed. To incorporate this scheme into the forecast recursion, we proceed as follows. In each step of the forecasting exercise, estimation is done by penalized least squares with the smoothing parameter selected by AIC. If the resulting forecast is negative, the PLS estimator is reobtained with the non-negativity constrained imposed as explained above, using the smoothing parameter that was selected for the original unconstrained PLS estimator. The forecast is then recomputed employing the constrained PLS estimator. To illustrate this issue, the percentage of negative forecasts produced by MIDAS-PLS is reported in the following analysis, given by the total number of negative (and then replaced) forecasts, divided by the total number of forecasts produced. These negative forecasts do not affect the forecast evaluation, however, as they do not enter relevant measures

of forecast accuracy.

In addition to evaluating the MIDAS approaches, it is of interest to assess the performance of the MIDAS models relative to alternative models of conditional volatility. To this end, the more traditional GARCH(1,1) model is studied as a benchmark, see Bollerslev (1986). The GARCH(1,1) represents a large class of models of conditional volatility. This model is chosen due to its simplicity and its good - albeit not best - performance relative to competing models from this class, see Brooks and Persand (2003). Moreover, Alper et al. (2012) also use the GARCH(1,1) model as a benchmark in their study of the parametric MIDAS approach. This GARCH(1,1) model is based on daily observations, and an analogous recursion applies to construct forecasts, using the most recent 1,000 daily observations for estimation. Here, an ARMA model is estimated to model the conditional mean of the return series in each step. Since the GARCH model uses daily observations, however, forecasts of volatility in the next period, that is, the next ten days or the next month, are constructed by summing up multi-step ahead forecasts of the daily variance over the relevant forecasting horizon.

Let $\{\widehat{\sigma}^2_{t+h,t}\}_{t=1}^{T_f}$ denote the series of forecasts formed at period t for horizon h, where T_f is the total number of forecasts obtained via the recursion described above. These are the forecasts from the MIDAS regressions or the GARCH model. We follow Raunig (2006) and evaluate forecasts according to the mean squared forecast error (MSFE)

$$\frac{1}{T_f}\sum_{t=1}^{T_f}\left(\sigma^2_{t+h,t}-\widehat{\sigma}^2_{t+h,t}\right)^2,$$

and the mean absolute forecast error (MAFE)

$$\frac{1}{T_f}\sum_{t=1}^{T_f}\left|\sigma^2_{t+h,t}-\widehat{\sigma}^2_{t+h,t}\right|,$$

Here, $\sigma^2_{t+h,t} = RV_{t+h,t}$ is the proxy for the true conditional variance. The MSFE imposes a larger penalty on large errors than the MAE.

1.4.6 Out-of-sample forecasting exercise: results

Table 1.2 shows MSFE and MAFE ratios of MIDAS-PLS relative to MIDAS-NLS and GARCH(1,1), respectively. Again, these ratios vary considerably, depending on the number of lags employed in the MIDAS regressions and the sampling scheme. Two main conlcusions can be drawn from this exercise, however. First, the performance of the non-parametric approach is

promising overall as it delivers more precise forecasts as measured by MSFE or MAFE. This gain in forecast accuracy is observed both relative to the alternative mixed frequency approach MIDAS-NLS and the GARCH(1,1) model based entirely on daily observations. The parametric estimation performs best relative to the non-parametric approach when $P = 40$, while penalized least squares offers more accurate forecasts when $P = 50$ or $P = 30$. Second, MIDAS-PLS is particularly advantageous for the monthly case as the ratios are smaller than with biweekly sampling. In addition, the number of negative (and thus meaningless) forecasts produced by MIDAS-PLS is small with about 5% of negative forecasts with biweekly sampling and at most 2% with monthly sampling. Taken together, the non-parametric MIDAS regression is a useful forecasting tool for medium to long-term volatility.

Table 1.2: Out-of-sample forecast comparison

		monthly sampling, $h = 1$			biweekly sampling, $h = 1$		
Sample: 1993-2013		$P = 50$	$P = 40$	$P = 30$	$P = 50$	$P = 40$	$P = 30$
MIDAS-NLS							
	MSFE	0.78	0.94	0.75	0.97	1.00	0.95
	MAFE	0.81	0.97	0.78	0.98	0.98	0.96
GARCH(1,1)							
	MSFE	0.86	0.95	0.85	0.95	0.96	0.93
	MAFE	0.95	0.93	0.93	0.98	0.94	0.94
	% negative	0.84	1.68	1.68	5.46	5.04	5.46

Note: Entries are MSFE and MAFE ratios of MIDAS-PLS relative to MIDAS-NLS and GARCH(1,1), respectively. The sample period has $T = 255$ ($T = 510$) monthly (biweekly) observations. Initial estimation sample from 1993-2003 with $T_e = 136$ ($T_e = 272$) monthly (biweekly) observations. Forecasts for MIDAS regressions are obtained with a rolling window of fixed size of T_e observations by successively dropping the first period in the sample and incorporating the following period. Total number of forecasts is $T_f = 119$ for the monthly and $T_f = 238$ for the biweekly case. The GARCH(1,1) forecast recursion uses a rolling estimation window of 1,000 days. The percentage of negative forecasts produced by MIDAS-PLS is reported, defined as the total number of negative forecasts divided by T_f. These negative forecasts do not enter the MSFE or MAFE ratios as they are replaced by the forecasts obtained from constrained PLS estimation in (1.25).

1.5 Concluding remarks

An alternative to parametric estimation of MIDAS regression models is suggested. The estimation procedure assumes that the true weighting function is smooth, but does not specify the functional form of an underlying weighting function explicitly. The estimator requires to specify a smoothing parameter and suitable methods to do so are presented.

The Monte Carlo experiment considers several empirically relevant weight functions and in these specifications, the non-parametric approach is able to compete with and has the potential to improve upon the parametric approach. Although in practice some experimentation with the degree of smoothing may be necessary, the modified AIC is a helpful guide. The non-parametric MIDAS regression is then used to forecast realized variance of the German stock index DAX at biweekly and monthly horizon, showing that the non-parametric MIDAS regression provides more precise forecasts then the parametric MIDAS with Beta lag structure or the GARCH(1,1) model using daily observations. In future research, this method may be useful either for the purpose of forecasting volatility by incorporating the non-parametric procedure into the GARCH-MIDAS model by Engle et al. (2013) or to produce macroeconomic forecasts via the ADL-MIDAS model of Andreou et al. (2013).

Appendix to Chapter 1

1.A First-order condition for minimizing AIC

In this appendix we derive the first-order condition for minimizing the AIC suggested by Hurvich, Simonoff, and Tsai (1998). Assuming a minimum exists in the interior of $\mathbb{R}_+$, we consider

$$\frac{d}{d\lambda}\text{AIC}(\lambda) = \frac{d}{d\lambda}\left(\log\left[\left(y^h - \widehat{y}_\lambda^h\right)'\left(y^h - \widehat{y}_\lambda^h\right)\right] + \frac{2(\kappa_\lambda + 1)}{T - \kappa_\lambda - 2}\right),$$

where $\kappa_\lambda = \text{tr}\left[X\left(X'X + \lambda D'D\right)^{-1}X'\right]$. First,

$$\frac{d}{d\lambda}\left(\log\left[\left(y^h - \widehat{y}_\lambda^h\right)'\left(y^h - \widehat{y}_\lambda^h\right)\right]\right) = \frac{1}{\text{SSR}_\lambda}\frac{d}{d\lambda}\left(\left(y^h - \widehat{y}_\lambda^h\right)'\left(y^h - \widehat{y}_\lambda^h\right)\right),$$

with $\text{SSR}_\lambda = \left(y^h - \widehat{y}_\lambda^h\right)'\left(y^h - \widehat{y}_\lambda^h\right)$. We establish two intermediate results: first,

$$\begin{aligned}
\frac{d}{d\lambda}Z_\lambda &= \frac{d}{d\lambda}\left(X\left(X'X + \lambda D'D\right)^{-1}X'\right) \\
&= X\left(\frac{d}{d\lambda}\left(X'X + \lambda D'D\right)^{-1}\right)X' \\
&= X\left(-\left(X'X + \lambda D'D\right)^{-1}\left(D'D\right)\left(X'X + \lambda D'D\right)^{-1}\right)X' \\
&= -XQ_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}X', \qquad (1.26)
\end{aligned}$$

with $Q_\lambda = (X'X + \lambda D'D)$. Second,

$$\begin{aligned}
\frac{d}{d\lambda}\left(Z_\lambda'Z_\lambda\right) &= \left(\frac{d}{d\lambda}\left(Z_\lambda'\right)\right)Z_\lambda + Z_\lambda'\left(\frac{d}{d\lambda}\left(Z_\lambda\right)\right) \\
&= \left(\frac{d}{d\lambda}\left(Z_\lambda\right)\right)'Z_\lambda + Z_\lambda'\left(\frac{d}{d\lambda}\left(Z_\lambda\right)\right) \\
&= \left(-XQ_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}X'\right)Z_\lambda + Z_\lambda'\left(-XQ_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}X'\right). \qquad (1.27)
\end{aligned}$$

Using (1.26) and (1.27),

$$\begin{aligned}
\frac{d}{d\lambda}\left(\left(y^h - \widehat{y}_\lambda^h\right)'\left(y^h - \widehat{y}_\lambda^h\right)\right) &= \frac{d}{d\lambda}\left(-2y^{h\prime}Z_\lambda y^h + y^{h\prime}Z_\lambda Z_\lambda' y^h\right) \\
&= -2y^{h\prime}\left(\frac{d}{d\lambda}\left(Z_\lambda\right)\right)y^h + y^{h\prime}\left(\frac{d}{d\lambda}\left(Z_\lambda'Z_\lambda\right)\right)y^h \\
&= 2y^{h\prime}XQ_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}X'\left(y^h - \widehat{y}_\lambda^h\right). \qquad (1.28)
\end{aligned}$$

Next,

$$\frac{d}{d\lambda}\left(\frac{2(\kappa_\lambda + 1)}{T - \kappa_\lambda - 2}\right) = \frac{2\frac{d}{d\lambda}(\kappa_\lambda)(T - \kappa_\lambda - 2) + 2(\kappa_\lambda + 1)\left(\frac{d}{d\lambda}(\kappa_\lambda)\right)}{(T - \kappa_\lambda - 2)^2} \qquad (1.29)$$

and

$$\begin{aligned}\frac{d}{d\lambda}\left(\kappa_\lambda\right) &= \frac{d}{d\lambda}\left(\text{tr}\left[Z_\lambda\right]\right) \\ &= \text{tr}\left[\frac{d}{d\lambda}\left(Z_\lambda\right)\right] \\ &= -\text{tr}\left[\left(X'X\right)Q_\lambda^{-1}\left(D'D\right)Q_\lambda\right]. \end{aligned} \tag{1.30}$$

Inserting (1.30) into (1.29) yields

$$\frac{d}{d\lambda}\left(\frac{2\left(\kappa_\lambda+1\right)}{T-\kappa_\lambda-2}\right) = -\frac{2\left(T-1\right)\text{tr}\left[\left(X'X\right)Q_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}\right]}{\left(T-\kappa_\lambda-2\right)^2}.$$

Taken together, a necessary condition for the minimizer $\lambda^\star_{\text{AIC}}$ is

$$\begin{aligned}\left(\text{SSR}_{\lambda^\star_{\text{AIC}}}\right)^{-1}\left(y^h{}'XQ^{-1}_{\lambda^\star_{\text{AIC}}}\left(D'D\right)Q^{-1}_{\lambda^\star_{\text{AIC}}}X'\left(y^h-\widehat{y}^h_{\lambda^\star_{\text{AIC}}}\right)\right) = \\ \frac{T-1}{\left(T-\kappa_{\lambda^\star_{\text{AIC}}}-2\right)^2}\text{tr}\left[\left(X'X\right)Q^{-1}_{\lambda^\star_{\text{AIC}}}\left(D'D\right)Q^{-1}_{\lambda^\star_{\text{AIC}}}\right].\end{aligned}$$

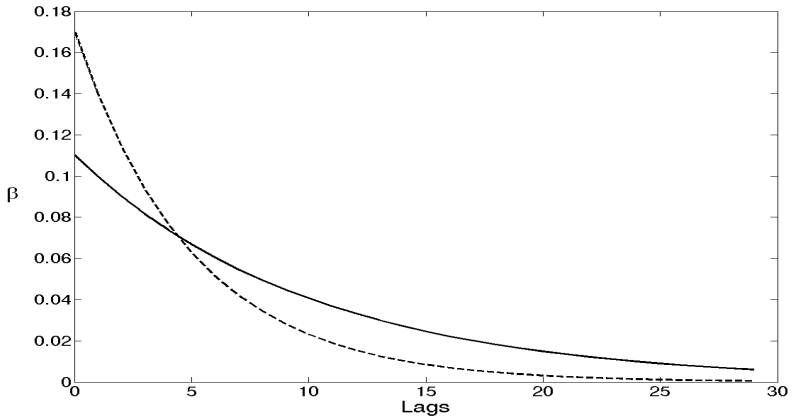
A sufficient condition for a minimizer is obtained by checking the second-order derivative at $\lambda^\star_{\text{AIC}}$. Using the same rules for matrix differential calculus as above, we find

$$\begin{aligned}&\frac{d^2}{d\lambda^2}\text{AIC}\left(\lambda\right) = \text{SSR}_\lambda^{-2}\left\{\left(2y^h{}'XQ_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}\left(X'X\right)Q_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}X'y^h-\right.\right. \\ &4y^h{}'XQ_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}X'\left(y^h-\widehat{y}^h_\lambda\right)\Big)\text{SSR}_\lambda- \\ &\left.\left(2y^h{}'XQ_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}X'\left(y^h-\widehat{y}^h_\lambda\right)\right)^2\right\}+4\left(T-1\right)\left(T-\kappa_\lambda-2\right)^{-3}\cdot \\ &\left\{\text{tr}\left[\left(X'X\right)Q_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}\right]\left(T-\kappa_\lambda-2\right)\right. \\ &\left.+\left(\text{tr}\left[\left(X'X\right)Q_\lambda^{-1}\left(D'D\right)Q_\lambda^{-1}\right]\right)^2\right\}.\end{aligned}$$

1.B Simulation results

Exponentially and linearly declining weights

Figure 1.2: Slow and moderate exponential decay



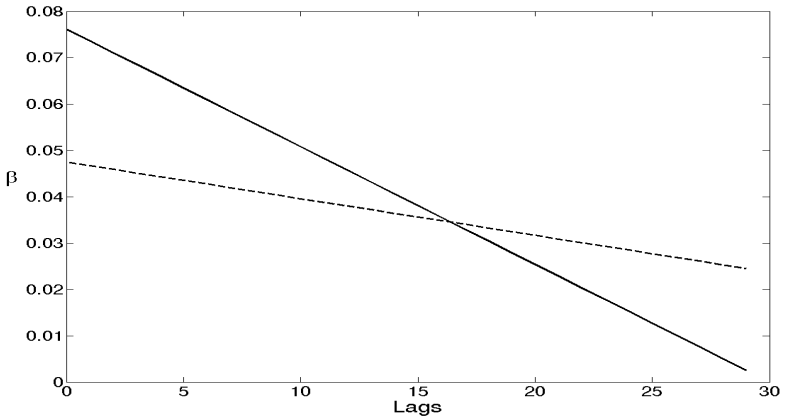
Note: Weight function according to (1.14) and (1.15) with $\theta_1 = -0.1$ (solid line) and $\theta_1 = -0.2$ (dashed line). The number of regressors is $P = 30$.

Table 1.3: MSE ratios for exponentially declining weight function

		$\theta_1 = -0.10$			$\theta_1 = -0.20$		
		$\psi = 0.8$	$\psi = 0.4$	$\psi = 0.2$	$\psi = 0.8$	$\psi = 0.4$	$\psi = 0.2$
$\mathbf{R^2 = 0.30}$		in-sample MSE ratios					
	$\overline{\lambda}_{\text{MSE}}$	0.70	0.85	0.86	0.93	1.26	1.30
	$\overline{\lambda}_{\text{AIC}}$	0.70	0.85	0.86	0.94	1.44	1.50
	$\overline{\lambda} = 2.5$	1.22	1.12	1.03	1.01	1.26	1.28
	$\overline{\lambda} = 5$	1.00	0.95	0.93	0.95	1.26	1.31
	$\overline{\lambda} = 10$	0.82	0.85	0.87	0.93	1.40	1.53
	$\overline{\lambda} = 20$	0.74	0.85	0.86	1.01	1.73	1.90
		out-of-sample MSFE ratios					
	$\overline{\lambda}_{\text{MSE}}$	0.98	0.99	0.99	0.99	0.99	0.99
	$\overline{\lambda}_{\text{AIC}}$	0.98	0.99	0.99	0.99	1.00	1.00
$\mathbf{R^2 = 0.15}$		in-sample MSE ratios					
	$\overline{\lambda}_{\text{MSE}}$	0.52	0.67	0.70	0.68	0.97	1.00
	$\overline{\lambda}_{\text{AIC}}$	0.57	0.67	0.69	0.74	0.99	1.02
	$\overline{\lambda} = 2.5$	1.13	1.04	0.96	0.93	1.09	1.07
	$\overline{\lambda} = 5$	0.92	0.87	0.85	0.79	1.00	1.00
	$\overline{\lambda} = 10$	0.73	0.76	0.75	0.70	0.97	1.00
	$\overline{\lambda} = 20$	0.63	0.72	0.69	0.67	1.06	1.12
		out-of-sample MSFE ratios					
	$\overline{\lambda}_{\text{MSE}}$	0.98	0.97	0.98	0.99	0.99	0.98
	$\overline{\lambda}_{\text{AIC}}$	0.99	0.97	0.98	0.99	0.99	0.99

Note: The sample size is $T = 100$ and number of regressors is $P = 30$. The smoothing parameter is reported relative to sample size $\overline{\lambda} = \lambda/T$.

Figure 1.3: Linearly declining and near-flat weights



Note: Weight function according to (1.14) and (1.16) with $a_0 = 1.5, a_1 = -0.05$ (solid line) and $a_0 = 3, a_1 = -0.05$ (dashed line). Sample size is $T = 100$ and number of regressors is $P = 30$.

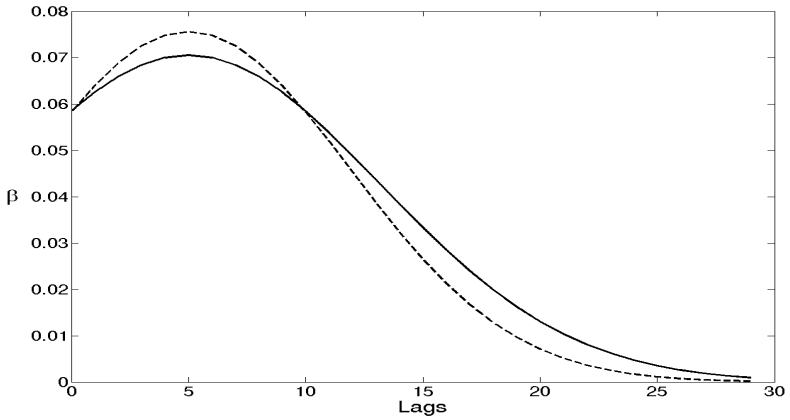
Table 1.4: MSE ratios for linearly declining weight function

		$a_0 = 1.5, a_1 = -0.05$			$a_0 = 3, a_1 = -0.05$		
		$\psi = 0.8$	$\psi = 0.4$	$\psi = 0.2$	$\psi = 0.8$	$\psi = 0.4$	$\psi = 0.2$
$\mathbf{R^2 = 0.30}$				in-sample MSE ratios			
	$\overline{\lambda}_{\text{MSE}}$	0.40	0.49	0.49	0.43	0.48	0.47
	$\overline{\lambda}_{\text{AIC}}$	0.60	0.60	0.57	0.64	0.59	0.53
	$\overline{\lambda} = 2.5$	1.39	1.15	1.09	1.48	1.12	1.03
	$\overline{\lambda} = 5$	1.11	1.00	0.95	1.19	0.97	0.90
	$\overline{\lambda} = 10$	0.89	0.84	0.78	0.95	0.82	0.74
	$\overline{\lambda} = 20$	0.77	0.70	0.66	0.82	0.68	0.63
				out-of-sample MSFE ratios			
	$\overline{\lambda}_{\text{MSE}}$	0.98	0.98	0.99	0.98	0.98	0.98
	$\overline{\lambda}_{\text{AIC}}$	0.99	0.98	0.99	0.99	0.98	0.99
$\mathbf{R^2 = 0.15}$				in-sample MSE ratios			
	$\overline{\lambda}_{\text{MSE}}$	0.37	0.49	0.48	0.39	0.46	0.44
	$\overline{\lambda}_{\text{AIC}}$	0.56	0.60	0.54	0.58	0.57	0.51
	$\overline{\lambda} = 2.5$	1.28	1.15	1.05	1.33	1.08	0.98
	$\overline{\lambda} = 5$	1.03	0.99	0.91	1.07	0.93	0.85
	$\overline{\lambda} = 10$	0.82	0.83	0.75	0.86	0.79	0.70
	$\overline{\lambda} = 20$	0.71	0.70	0.64	0.74	0.66	0.59
				out-of-sample MSFE ratios			
	$\overline{\lambda}_{\text{MSE}}$	0.97	0.97	0.98	0.98	0.97	0.97
	$\overline{\lambda}_{\text{AIC}}$	0.98	0.97	0.98	0.98	0.97	0.97

Note: For additional information, see the note in table 1.3.

Hump-shaped weights

Figure 1.4: Hump-shaped decay



Note: Weight function according to (1.14) and (1.17) with $\theta_1 = 0.105$, $\theta_2 = 0.0105$ (solid line) and $\theta_1 = 0.175$, $\theta_2 = 0.0175$ (dashed line). The number of regressors is $P = 30$.

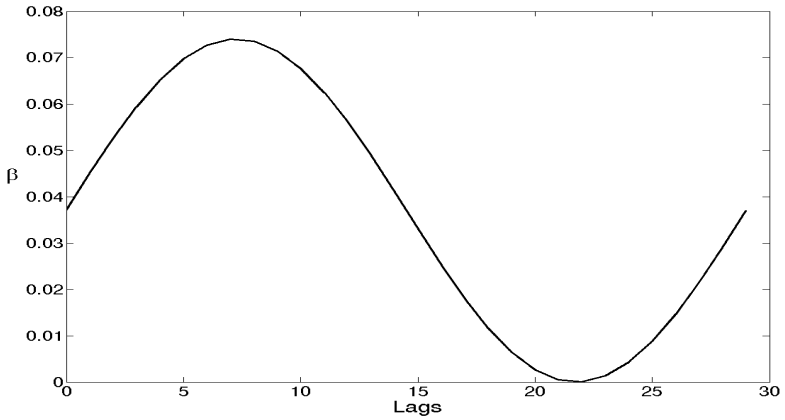
Table 1.5: MSE ratios for hump-shaped weight function

		$\theta_1 = 0.075, \theta_2 = 0.0075$			$\theta_1 = 0.105, \theta_2 = 0.0105$		
		$\psi = 0.8$	$\psi = 0.4$	$\psi = 0.2$	$\psi = 0.8$	$\psi = 0.4$	$\psi = 0.2$
$\mathbf{R^2 = 0.30}$		in-sample MSE ratios					
	$\overline{\lambda}_{\text{MSE}}$	0.69	0.97	0.99	0.79	1.28	1.26
	$\overline{\lambda}_{\text{AIC}}$	0.79	0.99	0.99	0.88	1.25	1.26
	$\overline{\lambda} = 2.5$	1.29	1.23	1.21	1.23	1.34	1.28
	$\overline{\lambda} = 5$	1.10	1.18	1.14	1.08	1.30	1.26
	$\overline{\lambda} = 10$	0.99	1.12	1.10	1.01	1.27	1.24
	$\overline{\lambda} = 20$	0.91	1.07	1.06	0.96	1.29	1.26
		out-of-sample MSFE ratios					
	$\overline{\lambda}_{\text{MSE}}$	0.98	1.00	1.00	0.98	0.99	1.00
	$\overline{\lambda}_{\text{AIC}}$	0.98	1.00	1.00	0.98	1.00	1.00
$\mathbf{R^2 = 0.15}$		in-sample MSE ratios					
	$\overline{\lambda}_{\text{MSE}}$	0.48	0.67	0.71	0.60	1.00	1.05
	$\overline{\lambda}_{\text{AIC}}$	0.61	0.73	0.74	0.68	1.01	1.05
	$\overline{\lambda} = 2.5$	1.18	1.15	1.10	1.00	1.19	1.18
	$\overline{\lambda} = 5$	0.98	1.03	1.02	0.89	1.10	1.13
	$\overline{\lambda} = 10$	0.83	0.93	0.91	0.81	1.06	1.10
	$\overline{\lambda} = 20$	0.73	0.83	0.83	0.75	1.03	1.07
		out-of-sample MSFE ratios					
	$\overline{\lambda}_{\text{MSE}}$	0.98	0.98	0.98	0.97	0.98	0.98
	$\overline{\lambda}_{\text{AIC}}$	0.99	0.98	0.98	0.98	0.98	0.98

Note: For additional information, see the note in table 1.3.

Cyclical weights

Figure 1.5: Cyclical weights



Note: Weight function according to (1.14) and (1.18). The number of regressors is $P = 30$.

Table 1.6: MSE ratios for cyclical weight function

		2-parameter Almon lag			3-parameter Almon lag		
		$\psi = 0.8$	$\psi = 0.4$	$\psi = 0.2$	$\psi = 0.8$	$\psi = 0.4$	$\psi = 0.2$
$\mathbf{R^2 = 0.30}$				in-sample MSE ratios			
	$\overline{\lambda}_{\text{MSE}}$	0.82	0.79	0.76	0.67	0.79	0.81
	$\overline{\lambda}_{\text{AIC}}$	0.93	1.00	1.00	0.76	0.99	1.05
	$\overline{\lambda} = 2.5$	0.84	0.82	0.84	0.69	0.81	0.88
	$\overline{\lambda} = 5$	0.85	0.96	1.02	0.69	0.95	1.08
	$\overline{\lambda} = 10$	0.95	1.15	1.21	0.77	1.14	1.28
	$\overline{\lambda} = 20$	1.09	1.30	1.34	0.89	1.29	1.41
				out-of-sample MSE ratios			
	$\overline{\lambda}_{\text{MSE}}$	0.99	0.99	0.98	0.98	0.98	0.98
	$\overline{\lambda}_{\text{AIC}}$	0.99	0.99	0.99	0.98	0.99	0.99
$\mathbf{R^2 = 0.15}$				in-sample MSE ratios			
	$\overline{\lambda}_{\text{MSE}}$	0.74	0.90	0.86	0.47	0.75	0.79
	$\overline{\lambda}_{\text{AIC}}$	0.79	0.93	0.92	0.49	0.77	0.85
	$\overline{\lambda} = 2.5$	0.95	0.90	0.87	0.60	0.75	0.79
	$\overline{\lambda} = 5$	0.83	0.90	0.89	0.52	0.75	0.82
	$\overline{\lambda} = 10$	0.81	0.93	0.95	0.51	0.78	0.87
	$\overline{\lambda} = 20$	0.82	0.99	0.97	0.52	0.82	0.89
				out-of-sample MSE ratios			
	$\overline{\lambda}_{\text{MSE}}$	0.98	0.99	0.99	0.97	0.98	0.97
	$\overline{\lambda}_{\text{AIC}}$	0.98	0.99	0.99	0.98	0.98	0.98

Note: For additional information, see the note in table 1.3.

1.C Summary statistics for the DAX index

Figure 1.6: DAX index

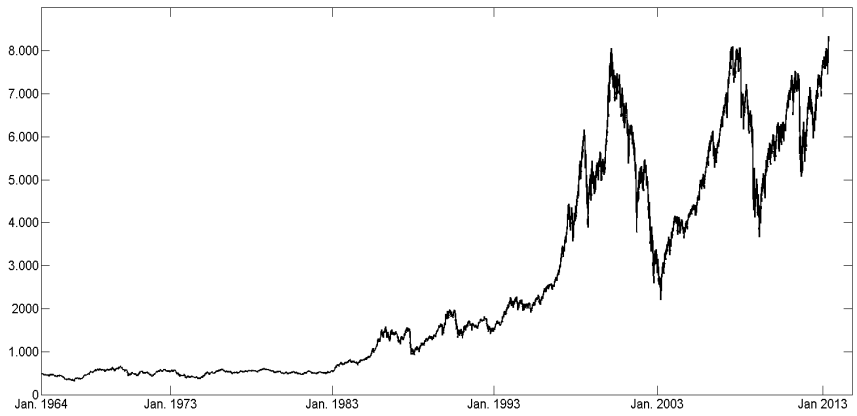


Figure 1.7: DAX log returns

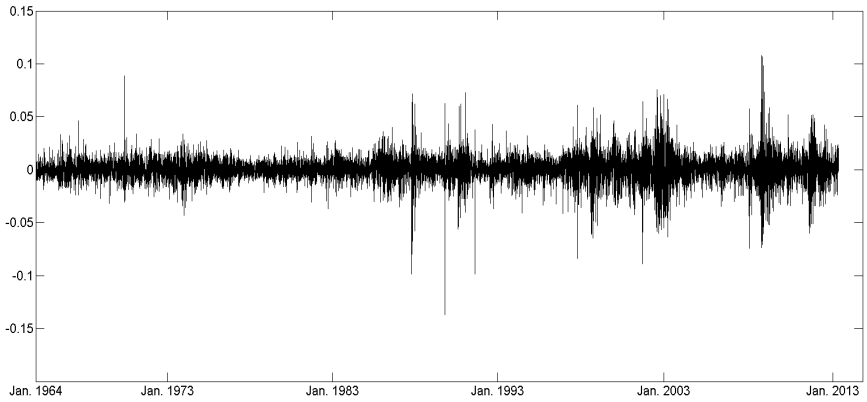


Table 1.7: Summary statistics for the DAX log return series

Sample	Mean	Std. dev.	Skewness	Kurtosis
Jan 1964 - May 2013	0.000234	0.012	-0.246	10.50
Jan 1993 - May 2013	0.000329	0.015	-0.129	7.41
Jan 2002 - May 2013	0.000165	0.016	0.054	7.62

Chapter 2

LM-type tests for slope homogeneity in panel data models

2.1 Introduction

In classical panel data analysis it is assumed that unobserved heterogeneity is captured by individual-specific constants, whether they are assumed to be fixed or random. In many applications, however, it cannot be ruled out that slope coefficients are also individual-specific. For instance, heterogenous preferences among consumers may imply individual-specific income elasticities. Ignoring this form of heterogeneity may result in biased estimation and inference. Therefore, it is important to test the assumption of slope homogeneity before applying standard panel data techniques such as the least-squares dummy-variable (LSDV) estimator for the fixed effect panel data model.

If there is evidence for individual-specific slope parameters, economists are interested in estimating a population average like the mean of the individual-specific coefficients. Pesaran and Smith (1995) advocate mean group estimation, where in a first step the model is estimated separately for each cross-section unit. In a second step, the unit specific estimates are averaged to obtain an estimator for the population mean of the parameters. Alternatively, Swamy (1970) proposes a generalized least squares (GLS) estimator for the random coefficients model, which assumes that the individual regression coefficients are randomly distributed around a common mean.

In this chapter, we take the random coefficients model as the starting point of our analysis. We derive the Lagrange Multiplier (LM) for the hypothesis that cross-sectional variance of the regression coefficients is equal to zero. To prepare the theoretical discussion in the follow-

ing sections, we briefly review the random coefficients model and a test suggested therein to investigate this issue. Following Swamy (1970), consider a linear panel data model

$$y_{it} = x_{it}'\beta_i + \epsilon_{it}, \tag{2.1}$$

for $i = 1, 2, \ldots, N$, and $t = 1, 2, \ldots, T$, where y_{it} is the dependent variable for unit i at time period t, x_{it} is a $K \times 1$ vector of explanatory variables and ϵ_{it} is an idiosyncratic error with zero mean and variance $\mathbb{E}\left[\epsilon_{it}^2\right] = \sigma_i^2$. The slope coefficients β_i are assumed to be randomly distributed with

$$\beta_i = \beta + v_i \ ,$$

where β is a fixed $K \times 1$ vector and v_i is a random vector with zero mean and $K \times K$ covariance matrix Σ_v.[1]

The null hypothesis of slope homogeneity is

$$\beta_1 = \beta_2 = \cdots = \beta_N = \beta, \tag{2.2}$$

which is equivalent to testing $\Sigma_v = 0$. To test hypothesis (2.2), Swamy suggests the statistic

$$\widehat{S}^* = \sum_{i=1}^{N} \left(\widehat{\beta}_i - \widehat{\beta}_{\text{WLS}}\right)' \left(\frac{X_i'X_i}{s_i^2}\right) \left(\widehat{\beta}_i - \widehat{\beta}_{\text{WLS}}\right),$$

with $X_i = (x_{i1}, \ldots, x_{iT})'$ and $\widehat{\beta}_i = (X_i'X_i)^{-1}X_i'y_i$ is the ordinary least squares (OLS) estimator of (2.1) for panel unit i, and $t = 1, \ldots, T$. The common slope parameter β is estimated by the weighted least-squares estimator

$$\widehat{\beta}_{\text{WLS}} = \left(\sum_{i=1}^{N} \frac{X_i'X_i}{s_i^2}\right)^{-1} \left(\sum_{i=1}^{N} \frac{X_i'y_i}{s_i^2}\right),$$

where s_i^2 denotes the standard OLS estimator of σ_i^2.

Intuitively, if the regression coefficients are identical, the differences between the individual estimators and the pooled estimator should be small. Therefore, Swamy's test rejects the null hypothesis of homogenous slopes for large values of this statistic, which possesses a limiting χ^2 distribution with $K(N-1)$ degrees of freedom as N is fixed and $T \to \infty$.

[1] For more details and extensions of the basic random coeffient model outlined here, see the thorough review by Hsiao and Pesaran (2008).

Pesaran and Yamagata (2008), henceforth referred to as PY, emphasize that in many empirical applications N is large relative to T and the approximation by a χ^2 distribution is unreliable. PY adapt the test to a setting in which N and T jointly tend to infinity. In particular, they assume individual-specific intercepts and derive a test for the hypothesis $\beta_1 = \cdots = \beta_N = \beta$ in

$$y_{it} = \alpha_i + x'_{it}\beta_i + \epsilon_{it}.$$

The analogue of the pooled weighted least squares estimator above eliminates the unobserved fixed effects,

$$\widehat{\beta}_{\text{WFE}} = \left(\sum_{i=1}^{N} \frac{X_i' M_\iota X_i}{\widehat{\sigma}_i^2} \right)^{-1} \left(\sum_{i=1}^{N} \frac{X_i' M_\iota y_i}{\widehat{\sigma}_i^2} \right),$$

where $M_\iota = I_T - \iota_T \iota_T'/T$, and ι_T is a $T \times 1$ vector of ones. A natural estimator for σ_i^2 is

$$\widehat{\sigma}_i^2 = \frac{\left(y_i - X_i\widehat{\beta}_i\right)' M_\iota \left(y_i - X_i\widehat{\beta}_i\right)}{T - K - 1},$$

where $\widehat{\beta}_i = \left(X_i' M_\iota X_i\right)^{-1} \left(X_i' M_\iota y_i\right)$ and the test statistic becomes

$$\widehat{S} = \sum_{i=1}^{N} \left(\widehat{\beta}_i - \widehat{\beta}_{\text{WFE}}\right)' \left(\frac{X_i' M_\iota X_i}{\widehat{\sigma}_i^2}\right) \left(\widehat{\beta}_i - \widehat{\beta}_{\text{WFE}}\right).$$

Employing a joint limit theory for N and T, PY obtain the limiting distribution as

$$\widehat{\Delta} = \frac{\widehat{S} - NK}{\sqrt{2NK}} \xrightarrow{d} \mathcal{N}(0,1), \tag{2.3}$$

provided that $N \to \infty$, $T \to \infty$ and $\sqrt{N}/T \to 0$. Thus, by appropriately centering and standardizing the test statistic, inference can be carried out by resorting to the standard normal distribution, provided the time dimension is sufficiently large relative to the cross-section dimension. PY propose several modified versions of this test, which for brevity we shall refer to as the Δ tests or statistics. In particular, to improve the small-sample properties of the test, PY suggest the adjusted statistic under normally distributed errors (see Remark 2 in PY),

$$\widetilde{\Delta}_{\text{adj}} = \sqrt{N(T+1)} \left(\frac{N^{-1}\widetilde{S} - K}{\sqrt{2K\,(T-K-1)}} \right), \tag{2.4}$$

where $\widetilde{S}$ is computed as $\widehat{S}$ but replacing $\widehat{\sigma}_i^2$ by the variance estimator

$$\widetilde{\sigma}_i^2 = \frac{\left(y_i - X_i\widetilde{\beta}_{\text{FE}}\right)' M_\iota \left(y_i - X_i\widetilde{\beta}_{\text{FE}}\right)}{T-1}, \tag{2.5}$$

where $\widetilde{\beta}_{\text{FE}} = \left(\sum_{i=1}^{N} X_i' M_\iota X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' M_\iota y_i\right)$ is the standard 'fixed effects' (within-group) estimator. Note that this asymptotic framework does not seem to be well suited for typical panel data applications where N is (very) large relative to T. Therefore, it will be of interest to derive a test statistic that is valid when T is small (say $T = 10$) and N is very large (say $N = 1000$), which, for instance, is encountered in microeconomic panels.

In this chapter we derive a test for slope homogeneity by employing the LM principle within a random coefficients framework, which allows us to formulate the null hypothesis of homogeneity in terms of K restrictions on the variance parameters of the parameter disturbances. Hence, the LM approach substantially reduces the number of restrictions to be tested compared to the set of $K(N-1)$ linear restrictions on the coefficients implied by the Swamy-PY approach. To derive the LM test, we assume that regression disturbances are normally distributed. We then provide variants of the LM statistic that robustify the original LM statistic to non-normally distributed and heteroskedastic errors. In addition, it is shown how the proposed LM statistics can be computed by running a simple artificial regression.

To gain further insight into the relationship between the LM and the Δ tests, we investigate the local asymptotic power of these tests in the random coefficient model. We find that the LM test and the Δ test have power against alternatives in a $N^{-1/2}T^{-1}$ neighborhood of the null hypothesis. The location parameter of the LM test depends on the cross-section dispersion of the regression variances, whereas the location parameter of the Δ test depends only on the mean of the regressor variances. Thus, if the regressor variances differ across the panel groups, the gain in power from using the LM test may be substantial.

It should be noted that in related work, Juhl and Lugovskyy (2013) derive a test for slope homogeneity in a likelihood framework with random coefficients. The resulting CLM test is based entirely on the score of their likelihood function, while the LM test proposed in this chapter also incorporates the information matrix, leading to a more powerful test.

To evaluate the performance of the LM-type tests in finite samples that are typically encountered in practice, we conduct several Monte Carlo experiments. The main conclusion in all experiments is that the LM test provides a sizeable power gain relative to existing procedures

when the time dimension is small. We emphasize that we restrict attention to static panels, as adopting the LM approach in a dynamic panel data model is beyond the scope of the chapter.
The outline of the chapter is as follows. In Section 2.2, we describe the random coefficients model and lay out the assumptions for analyzing the large-sample properties of our test statistics under normality. In Section 2.3 we derive the LM statistic and establish its asymptotic distribution. Section 2.4 discusses several variants of the proposed test. First, we relax the normality assumption and extend the result of the previous section to this more general setting. Second, we propose a regression-based version of the LM test. Section 2.5 investigates the local asymptotic power of the LM test. Section 2.6 describes the design of our Monte Carlo experiments and discusses the results. Section 2.7 concludes this chapter.

2.2 Model and assumptions

Consider a linear panel data model with random coefficients,

$$y_i = X_i\beta_i + \epsilon_i, \tag{2.6}$$

$$\beta_i = \beta + v_i, \tag{2.7}$$

for $i = 1, 2, \ldots, N$, where y_i is a is a $T \times 1$ vector of observations on the dependent variable for cross-section unit i, and X_i is a $T \times K$ matrix of possibly stochastic regressors. To simplify the exposition we assume a balanced panel with the same number of observation in each panel unit (see also Remark 2.1 of Lemma 2.1). The vector of random coefficients is assumed to have two components, the common non-stochastic vector β and a vector of individual-specific disturbances v_i.
Let $X = [X_1', X_2', \ldots, X_N']'$. We impose the following assumptions on the errors and the regressor matrix:

Assumption 2.1 *The error vectors are distributed as* $\epsilon_i|X \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I_T)$ *and* $v_i|X \overset{iid}{\sim} \mathcal{N}(0, \Sigma_v)$*, where* $\Sigma_v = diag\left(\sigma^2_{v,1}, \ldots, \sigma^2_{v,K}\right)$*. The errors* ϵ_i *and* v_j *are independent from each other for all* i *and* j*.*

Assumption 2.2 *For the regressors we assume* $\mathbb{E}|x_{it,k}|^{4+\delta} < C < \infty$ *for some* $\delta > 0$*, for all* $i = 1, 2 \ldots, N$*,* $t = 1, 2, \ldots, T$ *and* $k = 1, 2 \ldots, K$*. The limiting matrix* $\lim_{N\to\infty} N^{-1}\mathbb{E}\left[X'X\right]$ *exists and is positive definite for all* T*.*

In Assumption 2.1, the random components of the slope parameters are allowed to have different variances but we assume that there is no correlation among the elements of v_i. This assumption simplifies the derivation of the LM statistic. Allowing for correlation among the errors would increase the dimension of the null hypothesis to $K(K+1)/2$ restrictions and it is not clear whether including the covariances in the null hypothesis helps to increase the power of the test. Note that if all variances are zero, then the covariances are zero as well. We return to this issue when studying the small-sample properties of the LM test in Section 2.6.

To derive the asymptotic distribution of the LM statistic when $N \to \infty$ and T is fixed, we assume that errors are normally distributed. For the purpose of applying the LM test when errors are non-normally distributed, we generalize the setup in Section 2.4. The LM test is applicable in a framework in which $N, T \to \infty$ jointly once we trade off the distributional assumption for more specific restrictions on the existence of higher-order moments (see Theorem 2.2). Moreover, the regression error u_i is assumed to be homoskedastic. We propose a variant of the LM test that is robust to this assumption in Section 2.4.2.

Let $u_i = X_i v_i + \epsilon_i$. Stacking observations with respect to i yields

$$y = X\beta + u, \tag{2.8}$$

where $y = (y_1', \ldots, y_N')'$ and $u = (u_1', \ldots, u_N')'$. The $NT \times NT$ covariance matrix of u is given by

$$\Omega \equiv \mathbb{E}\left[uu'|X\right] = \begin{bmatrix} X_1\Sigma_v X_1' + \sigma^2 I_T & & 0 \\ & \ddots & \\ 0 & & X_N\Sigma_v X_N' + \sigma^2 I_T \end{bmatrix}.$$

The hypothesis of fixed homogeneous slope coefficients, $\beta_i = \beta$ for all i, corresponds to testing

$$H_0 : \ \sigma_{v,k}^2 = 0, \ \text{for } k = 1, ..., K,$$

against the alternative

$$H_1 : \ \sum_{k=1}^{K} \sigma_{v,k}^2 > 0, \tag{2.9}$$

that is, under the alternative at least one of the variance parameters is larger than zero.

2.3 The LM test for slope homogeneity

Let $\theta = \left(\sigma_{v,1}^2, ..., \sigma_{v,K}^2, \sigma^2\right)'$. Under Assumption 2.1 the corresponding log-likelihood function results as

$$\ell\left(\beta, \theta\right) = -\frac{NT}{2}\log(2\pi) - \frac{1}{2}\log\left|\Omega\left(\theta\right)\right| - \frac{1}{2}\left(y - X\beta\right)'\Omega\left(\theta\right)^{-1}\left(y - X\beta\right). \tag{2.10}$$

The restricted ML estimator of β under the null hypothesis coincides with the OLS estimator $\widetilde{\beta} = (X'X)^{-1}X'y$ in (2.8). Let $\widetilde{u} = y - X\widetilde{\beta}$ denote the corresponding vector of residuals. The restricted MLE of σ^2 is $\widetilde{\sigma}^2 = \widetilde{u}'\widetilde{u}/NT$.

The following lemma presents the score and the information matrix derived from the log-likelihood function in (2.10).

Lemma 2.1 *The score vector evaluated under the null hypothesis is given by*

$$\widetilde{\mathcal{S}} \equiv \left.\frac{\partial \ell}{\partial \theta}\right|_{H_0} = \frac{1}{2\widetilde{\sigma}^4}\begin{bmatrix} \sum_{i=1}^{N}\left(\widetilde{u}_i' X_i^{(1)} X_i^{(1)\prime}\widetilde{u}_i - \widetilde{\sigma}^2 X_i^{(1)\prime} X_i^{(1)}\right) \\ \vdots \\ \sum_{i=1}^{N}\left(\widetilde{u}_i' X_i^{(K)} X_i^{(K)\prime}\widetilde{u}_i - \widetilde{\sigma}^2 X_i^{(K)\prime} X_i^{(K)}\right) \\ 0 \end{bmatrix}, \tag{2.11}$$

where $X^{(k)}$ is the k-th column of X for $k = 1, 2, \ldots, K$.

The information matrix evaluated under the null hypothesis is

$$\begin{aligned} \mathcal{I}(\widetilde{\sigma}^2) \equiv & -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial\theta\partial\theta'}\right]\Bigg|_{H_0} \\ = & \frac{1}{2\widetilde{\sigma}^4}\begin{bmatrix} \sum_{i=1}^{N}\left(X_i^{(1)\prime} X_i^{(1)}\right)^2 & \cdots & \sum_{i=1}^{N}\left(X_i^{(1)\prime} X_i^{(K)}\right)^2 & X^{(1)\prime}X^{(1)} \\ \sum_{i=1}^{N}\left(X_i^{(2)\prime} X_i^{(1)}\right)^2 & \cdots & \sum_{i=1}^{N}\left(X_i^{(2)\prime} X_i^{(K)}\right)^2 & X^{(2)\prime}X^{(2)} \\ \vdots & \ddots & \vdots & \vdots \\ \sum_{i=1}^{N}\left(X_i^{(K)\prime} X_i^{(1)}\right)^2 & \cdots & \sum_{i=1}^{N}\left(X_i^{(K)\prime} X_i^{(K)}\right)^2 & X^{(K)\prime}X^{(K)} \\ X^{(1)\prime}X^{(1)} & \cdots & X^{(K)\prime}X^{(K)} & NT \end{bmatrix}, \end{aligned} \tag{2.12}$$

where $X_i^{(k)}$ denotes the k-th column of the $T \times K$ matrix X_i, $k = 1, 2, \ldots, K$ and $i = 1, ..., N$.

Remark 2.1 It is straightforward to extend Lemma 2.1 to unbalanced panel data. Assume that X_i is a $T_i \times K$ matrix and $\widetilde{u}_i$ is a conformable $T_i \times 1$ vector. The score vector is given by

$$\widetilde{\mathcal{S}} = \frac{1}{2\widetilde{\sigma}^4} \begin{bmatrix} \sum_{i=1}^{N} \left(\widetilde{u}_i' X_i^{(1)} X_i^{(1)\prime} \widetilde{u}_i - \widetilde{\sigma}^2 X_i^{(1)\prime} X_i^{(1)} \right) \\ \vdots \\ \sum_{i=1}^{N} \left(\widetilde{u}_i' X_i^{(K)} X_i^{(K)\prime} \widetilde{u}_i - \widetilde{\sigma}^2 X_i^{(K)\prime} X_i^{(K)} \right) \\ 0 \end{bmatrix},$$

where

$$\widetilde{\sigma}^2 = \frac{1}{\sum_{i=1}^{N} T_i} \sum_{i=1}^{N} \widetilde{u}_i' \widetilde{u}_i \, .$$

The information matrix is computed accordingly.

In the following theorem it is shown that when T is fixed the LM statistic possesses a χ^2 limiting null distribution with K degrees of freedom as $N \to \infty$.

Theorem 2.1 *Under Assumptions 2.1, 2.2 and the null hypothesis*

$$LM = \widetilde{\mathcal{S}}' \mathcal{I}(\widetilde{\sigma}^2)^{-1} \widetilde{\mathcal{S}} = \widetilde{s}' \, \widetilde{V}^{-1} \, \widetilde{s} \xrightarrow{d} \chi^2_K, \tag{2.13}$$

as $N \to \infty$ and T is fixed, where $\widetilde{s}$ is defined as the $K \times 1$ vector with typical element

$$\widetilde{s}_k = \frac{1}{2\widetilde{\sigma}^4} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \widetilde{u}_{it} x_{it,k} \right)^2 - \frac{1}{2\widetilde{\sigma}^2} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it,k}^2, \tag{2.14}$$

and the (k,l) element of the matrix $\widetilde{V}$ is given by

$$\widetilde{V}_{k,l} = \frac{1}{2\widetilde{\sigma}^4} \left[\sum_{i=1}^{N} \left(\sum_{t=1}^{T} x_{it,k} x_{it,l} \right)^2 - \frac{1}{NT} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} x_{it,k}^2 \right) \left(\sum_{i=1}^{N} \sum_{t=1}^{T} x_{it,l}^2 \right) \right]. \tag{2.15}$$

Remark 2.2 If T is fixed, normality of the regression disturbances is required. If we relax the normality assumption, an additional term enters the variance of the score vector and the information matrix becomes an inconsistent estimator. Theorem 2.2 discusses this issue in more details and derives the asymptotic distribution of the LM test if the errors are not normally distributed.

Remark 2.3 It may be of interest to restrict attention to a subset of coefficients. For example, in the classical panel data model it is assumed that the constants are individual-specific and,

therefore, the respective parameters are not included in the null hypothesis. Another possibility is that a subset of coefficients is assumed to be constant across all panel units. To account for such specifications the model is partitioned as

$$y_{it} = \beta_{1i}' X_{it}^a + \beta_2' X_{it}^b + \beta_{3i}' X_{it}^c + u_{it} \; .$$

The $K_1 \times 1$ vector X_{it}^a includes all regressors that are assumed to have individual-specific coefficients stacked in the vector β_{1i}. The $K_2 \times 1$ vector X_{it}^b comprises all regressors that are supposed to have homogenous coefficients. The null hypothesis is that the coefficient vector β_{3i} attached to the $K_3 \times 1$ vector of regressors X_{it}^c is identical for all panel units, that is, $\beta_{3i} = \beta_3$ for all i, where $\beta_{3i} = \beta_3 + v_{3i}$. The null hypothesis implies $\Sigma_{v_3} = 0$. Let

$$Z = \begin{bmatrix} X_1^a & 0 & \cdots & 0 & X_1^b & X_1^c \\ 0 & X_2^a & \cdots & 0 & X_2^b & X_2^c \\ \vdots & & \ddots & \vdots & & \\ 0 & 0 & \cdots & X_N^a & X_N^b & X_N^c \end{bmatrix},$$

where $X_i^a = [X_{i1}^a, \ldots, X_{iT}^a]'$ and the matrices X_i^b and X_i^c are defined accordingly. The residuals are obtained as $\widetilde{u} = (I - Z(Z'Z)^{-1}Z')y$ and the columns of the matrix X^c are used to compute the LM statistic. Some caution is required if a set of individual-specific coefficients are included in the panel regression since in this case the ML estimator $\widetilde{\sigma}^2 = (NT)^{-1} \sum_{i=1}^N \sum_{t=1} \widetilde{u}_{it}^2$ is inconsistent for fixed T and $N \to \infty$. This implies that the expectation of the score vector (2.11) is different from zero. Accordingly, the unbiased estimator

$$\widehat{\sigma}^2 = \frac{1}{NT - K_1 - K_2 - K_3} \sum_{i=1}^N \sum_{t=1}^T \widetilde{u}_{it}^2 \tag{2.16}$$

must be employed.

As a special case, assume that the constant is included in X_i^c, whereas all other regressors are included in the matrix X_i^b, and X_i^a is dropped. This case is equivalent to the test for random individual effects as suggested by Breusch and Pagan (1980). The LM statistic then reduces to

$$LM = \frac{NT}{2\,(T-1)} \left[1 - \frac{\widetilde{u}'\,(I_N \otimes \iota_T \iota_T')\,\widetilde{u}}{\widetilde{u}'\widetilde{u}} \right]^2 ,$$

where ι_T is a $T \times 1$ vector of ones, which is identical to the familiar LM statistic for random individual effects.

Remark 2.4 Let $\widetilde{s}_i$ be the $K \times 1$ vector of score contributions of unit i, such that the vector $\widetilde{s}$ defined in theorem 2.1 can be decomposed as $\widetilde{s} = \sum_{i=1}^{N} \widetilde{s}_i$. By replacing the matrix $\widetilde{V}$ in (2.13) by the "Outer Product of Gradients" (OPG), the OPG variant of the LM test results as

$$\left(\sum_{i=1}^{N} \widetilde{s}_i\right)' \left(\sum_{i=1}^{N} \widetilde{s}_i \widetilde{s}_i'\right)^{-1} \left(\sum_{i=1}^{N} \widetilde{s}_i\right).$$

with

$$\widetilde{s}_i = \frac{1}{2\widetilde{\sigma}^4} \left(\widetilde{u}_i' X_i X_i' \widetilde{u}_i - \widetilde{\sigma}^2 X_i' X_i\right). \tag{2.17}$$

This statistic can be related to the CLM statistic proposed by Juhl and Lugovskyy (2013). Their likelihood framework takes the fixed effects model as given and assumes that the regression errors are heterokedastic. Moreover, a random coefficients model is adopted with $\Sigma_v = \sigma_v^2 I_K$ (cf. Assumption 2.1). The general expression for the score contributions in their model is

$$\widetilde{s}_i^{\text{clm}} = \left(y_i - X_i \widetilde{\beta}_{\text{FE}}\right)' M_\iota X_i X_i' M_\iota \left(y_i - X_i \widetilde{\beta}_{\text{FE}}\right) - \widetilde{\sigma}_i^2 \text{tr}\left[X_i' M_\iota X_i\right], \tag{2.18}$$

where tr $[\cdot]$ denotes the trace of a matrix and $\widetilde{\sigma}_i^2$ is given in (2.5). The resulting test statistic is

$$CLM = \frac{\sum_{i=1}^{N} \widetilde{s}_i^{clm}}{\left(\sum_{i=1}^{N} \left(\widetilde{s}_i^{clm}\right)^2\right)^{1/2}}.$$

Comparing (2.17) and (2.18), for $K = 1$, the OPG variant of the LM statistic as given above can be viewed as the analogue of the CLM statistic in absence of fixed effects and with homoskedastic errors. In general, as shown in theorem 4.2 in Juhl and Lugovskyy (2013), under the associated null hypothesis $\sigma_v^2 = 0$,

$$CLM \xrightarrow{d} \mathcal{N}(0, 1), \tag{2.19}$$

as $N \to \infty$ and T is fixed.

2.4 Variants of the LM test

2.4.1 The LM statistic under non-normality

In this section we consider useful variants of the original LM statistic. First, we analyze the LM test under the assumption that the errors are not normally distributed. Therefore, we alter Assumptions 2.1 and 2.2 as follows.

Assumption 2.3 *The error vector ϵ_i is independently and identically distributed with $E(\epsilon_i|X) = 0$, $E(\epsilon_i\epsilon_i'|X) = \sigma^2 I_T$ and $E|\epsilon_{it}|^6 < C < \infty$ for all i and t. ϵ_i and v_j are independent from each other for all i and j.*

Assumption 2.4 *For the regressors we assume $\mathbb{E}|x_{it,k}|^{4+\delta} < C < \infty$ for some $\delta > 0$, for all $i = 1, 2 \ldots, N$, $t = 1, 2, \ldots, T$ and $k = 1, 2 \ldots, K$. The limiting matrix*

$$Q := \lim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[x_{it} x_{it}'\right]$$

exists and is positive definite.

With these modifications of the previous setup, the limiting distribution of the LM statistic is χ^2 distributed as $N, T \to \infty$ jointly.

Theorem 2.2 *Under Assumptions 2.3, 2.4 and the null hypothesis,*

$$LM \xrightarrow{d} \chi^2_K, \tag{2.20}$$

and $N \to \infty$, $T \to \infty$ jointly.

Generalizing the model to allow for non-normally distributed errors introduces a new term into the variance of the score: the (k, l) element of the covariance matrix now becomes (see equation (2.44) in appendix 2.A)

$$V_{k,l} + \left(\frac{\mu_u^{(4)} - 3\sigma^4}{\left(2\sigma^4\right)^2}\right) \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it,k}^2 - \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it,k}^2\right) \left(x_{it,l}^2 - \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it,l}^2\right), \tag{2.21}$$

where $\mu_u^{(4)}$ denotes the fourth moment of the error distribution, and $V_{k,l}$ is as in (2.15) with $\widetilde{\sigma}^4$ replaced by σ^4. The additional term depends on the excess kurtosis $\mu_u^{(4)} - 3\sigma^4$. Clearly, for

normally distributed errors, this term disappears, but it deviates from zero in the more general setup. Under Assumptions 2.3 and 2.4, the first term $V_{k,l}$ is of order NT^2, while the new component is of order NT, such that, when the appropriate scaling underlying the LM statistic is adopted, it vanishes as $T \to \infty$. Therefore, the LM statistic as presented in the previous section continues to be χ^2_K distributed asymptotically.

By incorporating a suitable estimator of the second term in (2.21), however, a test statistic becomes available that is valid in a framework with non-normally distributed errors as $N \to \infty$, whether T is fixed or $T \to \infty$. Therefore, denote the adjusted LM statistic by

$$LM_{\text{adj}} = \widetilde{s}' \left(\widetilde{V}_{\text{adj}} \right)^{-1} \widetilde{s},$$

where $\widetilde{V}_{\text{adj}}$ is as in (2.21) with $V_{k,l}$, σ^4 and $\mu_u^{(4)}$ replaced by the consistent estimators $\widetilde{V}_{k,l}$ for $k, l = 1, ..., K$, defined in (2.15), $\widetilde{\sigma}^4$ and $\widetilde{\mu}_u^{(4)} = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{u}_{it}^4$.

As a consequence of Theorem 2.2 and the preceding discussion, we obtain the following result.

Corollary 2.1 *Under Assumptions 2.2, 2.3 and the null hypothesis*

$$LM_{\text{adj}} \xrightarrow{d} \chi^2_K,$$

as $N \to \infty$ and T is fixed. Furthermore,

$$LM_{\text{adj}} - LM \xrightarrow{p} 0,$$

as $N \to \infty$, $T \to \infty$ jointly.

As mentioned above, once the regression disturbances are no longer normally distributed, the fourth moments of the error distribution enter the variance of the score. It is insightful to identify exactly which terms give rise to this new form of the covariance matrix. According to Lemma 2.1, the contribution of the i-th panel unit to the k-th element of the score vector is

$$\widetilde{u}_i' X_i^{(k)} X_i^{(k)\prime} \widetilde{u}_i - \widetilde{\sigma}^2 X_i^{(k)\prime} X_i^{(k)} = \left(\sum_{t=1}^{T} x_{it,k}^2 \left(\widetilde{u}_{it}^2 - \widetilde{\sigma}^2 \right) \right) + \sum_{t=1}^{T} \sum_{s \neq t} \widetilde{u}_{it} \widetilde{u}_{is} x_{it,k} x_{is,k}. \quad (2.22)$$

The first term has expectation zero whether the null hypothesis is true or not. Moreover, the variance of this term introduces the fourth moments of the errors into the variance of the score. Given this observation, the first term can be dropped from the analysis without affecting the asymptotic size or power of the test. We can then proceed to examine the asymptotic properties

of the remaining term when T is fixed. Hence, we consider a modified score vector as presented in the following theorem.

Theorem 2.3 *Under Assumptions 2.2, 2.3 and the null hypothesis, the modified LM statistic*

$$LM^* = \tilde{s}^{*\prime} \left(\widetilde{V}^*\right)^{-1} \tilde{s}^* \xrightarrow{d} \chi^2_K,$$

as $N \to \infty$ and T fixed, where $\tilde{s}^*$ is $K \times 1$ vector with contributions for panel unit i

$$\tilde{s}^*_{i,k} = \frac{1}{\widetilde{\sigma}^4} \sum_{t=2}^{T} \sum_{s=1}^{t-1} \widetilde{u}_{it} \widetilde{u}_{is} x_{it,k} x_{is,k}, \tag{2.23}$$

for $i = 1, ..., N$, $k = 1, ..., K$, and the (k, l) element of $\widetilde{V}^*$ is given by

$$\widetilde{V}^*_{k,l} = \frac{1}{\widetilde{\sigma}^4} \sum_{i=1}^{N} \sum_{t=2}^{T} x_{it,k} x_{it,l} \left(\sum_{s=1}^{t-1} x_{is,k} x_{is,l} \right), \tag{2.24}$$

for $k, l = 1, ..., K$.

Remark 2.5 It is important to note that this version of the LM test is invalid if the panel regression allows for individual-specific coefficients (cf. Remark 3). Consider for example the regression

$$y_{it} = \mu_i + x'_{it}\beta_i + u_{it}, \tag{2.25}$$

where μ_i are fixed individual effects and we are interested in testing $H_0 : Var[\beta_i] = 0$. The residuals are obtained as

$$\widetilde{u}_{it} \;=\; y_{it} - \overline{y}_i - (x_{it} - \overline{x}_i)'\widetilde{\beta} \;=\; u_{it} - \overline{u}_i - (x_{it} - \overline{x}_i)'(\widetilde{\beta} - \beta).$$

It follows that in this case $E(\widetilde{u}_{it}\widetilde{u}_{is}x_{it,k}x_{is,k}) \neq 0$ and, therefore, the modified scores (2.23) result in a biased test. To sidestep this difficulty, orthogonal deviations (e.g. Arellano and Bover (1995)) can be employed to eliminate the individual-specific constants yielding

$$y^*_{it} = \beta' x^*_{it} + u^*_{it} \quad t = 2, 3, \ldots, T,$$
$$\text{with } y^*_{it} = \sqrt{\frac{t-1}{t}} \left[y_{it} - \frac{1}{t-1} \left(\sum_{s=1}^{t-1} y_{is} \right) \right],$$

where x^*_{it} and u^*_{it} are defined analogously. It is well known that if u_{it} is serially uncorrelated so is u^*_{it}. It follows that the modified LM statistic can be constructed by using the OLS residuals $\widetilde{u}^*_{it}$ instead of $\widetilde{u}_{it}$. This approach can be generalized to arbitrary individual-specific regressors x^a_{it}. Let $X^a_i = [x^a_{i1}, \ldots, x^a_{iT}]'$ denote the individual-specific $T \times K_1$ regressor matrix in the regression

$$y_i = X^a_i \beta_{1i} + X^b_i \beta_2 + X^c_i \beta_{3i} + u_i, \tag{2.26}$$

(see Remark 3). Furthermore, let

$$M^a_i \;=\; I_T - X^a_i (X^{a\prime}_i X^a_i)^{-1} X^{a\prime}_i,$$

and let $\widetilde{M}^a_i$ denote the $(T - K_1) \times T$ matrix that results from eliminating the last K_1 rows from M^a_i such that $(M^a_i M^{a\prime}_i)$ is of full rank. The model (2.26) is transformed as

$$y^*_i = X^{b*}_i \beta_2 + X^{c*}_i \beta_{3i} + u^*_i \;, \tag{2.27}$$

where $y^*_i = \Xi^a_i y_i$ and $\Xi^a_i = (\widetilde{M}^a_i \widetilde{M}^{a\prime}_i)^{-1/2} \widetilde{M}^a_i$. It is not difficult to see that $E(u^*_i u^{*\prime}_i) = \sigma^2 I_{T-K_1}$ and, thus, the modified scores (2.23) can be constructed by using the residuals of (2.27), where the time series dimension reduces to $T - K_1$. Note that orthogonal deviations result from letting X^a_i be a vector of ones.

To review the results of this section, the important new feature in the model without assuming normality is that the fourth moments of the errors enter the variance of the score. The information matrix of the original LM test derived under normality does not incorporate higher order moments, but the test remains applicable as $T \to \infty$. To apply the LM test in the original framework when T is fixed and errors are no longer normal we can proceed in two ways. A direct adjustment of the information matrix to account for higher order moments yields a valid test. Alternatively, we can adjust the score itself and restrict attention to that part of the score that does not introduce higher order moments into the variance. In the next section, we further pursue the second route of dealing with non-normality and thereby robustify the test against heteroskedasticity.

2.4.2 The regression-based LM statistic

In this section we offer a convenient way to compute the proposed LM statistic via a simple artificial regression. Moreover, the regression-based form of the LM test is shown to be robust against heteroskedastic errors.

Following the decomposition of the score contribution in (2.22) and the discussion thereafter, we construct the "Outer Product of Gradients" (OPG) variant of the LM test based on the second term in (2.22). Rewriting the corresponding elements of the score contributions of panel unit i as

$$\tilde{s}_{i,k}^{*} = \frac{1}{2\widetilde{\sigma}^4} \sum_{t=1}^{T} \sum_{s \neq t} \widetilde{u}_{it} \widetilde{u}_{is} x_{it,k} x_{is,k} = \frac{1}{\widetilde{\sigma}^4} \sum_{t=2}^{T} \sum_{s=1}^{t-1} \widetilde{u}_{it} \widetilde{u}_{is} x_{it,k} x_{is,k}, \tag{2.28}$$

for $k = 1, ..., K$, gives the LM-OPG variant

$$LM_{\text{opg}} = \left(\sum_{i=1}^{N} \tilde{s}_i^{*} \right)' \left(\sum_{i=1}^{N} \tilde{s}_i^{*} \tilde{s}_i^{*\prime} \right)^{-1} \left(\sum_{i=1}^{N} \tilde{s}_i^{*} \right), \tag{2.29}$$

where $\tilde{s}_i^{*} = \left[\tilde{s}_{i,1}^{*}, ..., \tilde{s}_{i,K}^{*} \right]'$. An asymptotically equivalent form of the LM-OPG statistic can be formulated as a Wald-type test for the null hypothesis $\varphi = 0$ in the auxiliary regression

$$\widetilde{u}^{*} = \widetilde{Z} \varphi + e,$$

where $\widetilde{u}^{*} = (\widetilde{u}_{12}, \ldots, \widetilde{u}_{1T}, \ldots, \widetilde{u}_{N2}, \ldots, \widetilde{u}_{NT})'$, $\widetilde{Z} = \left[\widetilde{Z}_1', \ldots, \widetilde{Z}_N' \right]'$, and $\widetilde{Z}_i$ is a $(T-1) \times K$ matrix with typical element

$$\widetilde{z}_{it,k} = \frac{1}{\widetilde{\sigma}^4} x_{it,k} \sum_{s=1}^{t-1} \widetilde{u}_{is} x_{is,k},$$

for $k = 1, ..., K$ and $t = 2, ..., T$. Therefore, with the Eicker-White heteroskedasticity consistent variance estimator, the regression based test statistic results as

$$LM_{\text{reg}} = \left(\widetilde{Z}' \widetilde{u}^{*} \right)' \left(\sum_{i=1}^{N} \sum_{t=2}^{T} \widetilde{u}_{it}^2 \widetilde{z}_{it} \widetilde{z}_{it}' \right)^{-1} \left(\widetilde{Z}' \widetilde{u}^{*} \right), \tag{2.30}$$

where $\widetilde{z}_{it} = \left[\widetilde{z}_{it,1}, \ldots, \widetilde{z}_{it,K} \right]'$. Since $\sum_{i=1}^{N} \tilde{s}_i^{*} = \widetilde{Z}' \widetilde{u}^{*}$ and the variance estimators in (2.29) and the middle matrix in (2.30) are asymptotically equivalent, we obtain the following theorem.

Theorem 2.4 *Under Assumption 2.2, Assumption 2.3 but allowing for arbitrary variances* $\mathbb{E}(\epsilon_{it}^2) = \sigma_i^2 < \infty$*, and the null hypothesis, it holds that*

$$LM_{\text{reg}} \xrightarrow{d} \chi_K^2,$$

as $N \to \infty$ and T is fixed.

This result ends our discussion of examining the consequences of non-normal and heteroskedastic errors for our test. We have generalized the classical assumptions in Section 2.3 to settings that are commonly encountered in practice, including non-normal regression errors or heteroskedastic disturbances. By extending and robustifying the original LM test in these directions, we have now assembled a collection of test statistics to choose from. To study the power of the LM test, the following section examines the power of the class of tests against a suitable sequence of local alternatives.

2.5 Local power analysis

The aim of this section is twofold. First, we investigate the distributions of the LM-type test under suitable sequences of local alternatives. Two cases are of interest, with T fixed and $N, T \to \infty$ jointly, which are presented in the respective theorems below. Second, we adopt the results of PY to our model in order to compare the local asymptotic power of the two tests.
To formulate an appropriate sequence of local alternatives, we specify the random coefficients in (2.7) in a setup in which T is fixed. The error term v_i is as in Assumption 2.1 with elements of Σ_v given by

$$\sigma_{v,k}^2 = \frac{c_k}{\sqrt{N}}, \tag{2.31}$$

where $c_k > 0$ are fixed constants for $k = 1, \ldots, K$. The asymptotic distribution of the LM statistic results then as follows.

Theorem 2.5 *Under Assumptions 2.1, 2.2 and the sequence of local alternatives (2.31),*

$$LM \xrightarrow{d} \chi_K^2(\mu),$$

as $N \to \infty$ and T fixed, with non-centrality parameter $\mu = c'\Psi c$, where $c = (c_1, \ldots, c_K)'$ and Ψ is a $K \times K$ matrix with (k, l) element

$$\Psi_{k,l} = \frac{1}{2\sigma^4} \operatorname*{plim}_{N\to\infty} \left[\frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} x_{it,k} x_{it,l} \right)^2 - \frac{1}{T} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it,k}^2 \right) \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it,l}^2 \right) \right].$$

Similar conclusions hold if we relax normality and adopt Assumption 2.3 for v_i, where the

sequence of local alternatives is now given by

$$\sigma_{v,k}^2 = \frac{c_k}{T\sqrt{N}}, \tag{2.32}$$

for $k = 1, \ldots, K$.

Theorem 2.6 *Under Assumptions 2.3, 2.4 and the sequence of alternatives (2.32),*

$$LM \xrightarrow{d} \chi_K^2(\mu),$$

as $N \to \infty$, $T \to \infty$, with non-centrality parameter $\mu = c'\Psi c$, where $c = (c_1, \ldots, c_K)'$ and Ψ is a $K \times K$ matrix with (k, l) element

$$\Psi_{k,l} = \frac{1}{2\sigma^4} \operatorname*{plim}_{N,T\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{T} \sum_{t=1}^{T} x_{it,k} x_{it,l} \right)^2.$$

Remark 2.6 As in Section 2.4.1 above, when the normality assumption is relaxed, local power can be studied for LM^* under Assumptions 2.2 and 2.3 when T is fixed. The specification of local alternatives as in Theorem 2.5 applies. The non-centrality parameter of the limiting non-central χ^2 distribution results as $\mu^* = c'\Psi^* c$ with

$$\Psi_{k,l}^* = \frac{1}{\sigma^4} \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T} x_{it,k} x_{it,l} \left(\sum_{s=1}^{t-1} x_{is,k} x_{is,l} \right),$$

for $k, l = 1, ..., K$.

Remark 2.7 Given the results for the score-modified statistic LM^* in remark 2.6, and the fact that $\tilde{s}^* = \sum_{i=1}^{N} \tilde{s}_i^* = \left(\widetilde{Z}'\widetilde{u}^* \right)$, we expect a similar result for the regression-based LM statistic LM_{reg} to hold. Recall that LM^* uses $N^{-1}\widetilde{V}^*$ as an estimator of the variance of $\tilde{s}^*$ (see (2.24)), while LM_{reg} employs $\left(N^{-1} \sum_{i=1}^{N} \sum_{t=2}^{T} \widetilde{u}_{it}^2 \widetilde{z}_{it} \widetilde{z}_{it}' \right)$. Under the null hypothesis, it is not difficult to see that these two estimators are asymptotically equivalent, see the proof of Theorem 2.4. Under the alternative, when studying the (k, l) element of the variance of LM_{reg}, we obtain (see appendix 2.A for details)

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T} \widetilde{u}_{it}^2 \widetilde{z}_{it,k} \widetilde{z}_{it,l} = & \left(\frac{1}{\widetilde{\sigma}^4} \right)^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T} \epsilon_{it}^2 x_{it,k} x_{it,l} \left(\sum_{s=1}^{t-1} \epsilon_{is} x_{is,k} \right) \left(\sum_{s=1}^{t-1} \epsilon_{is} x_{is,l} \right) \\ & + \frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T} \epsilon_{it}^2 v_i' B_{it}^X v_i + o_p(1), \end{aligned} \tag{2.33}$$

with the $K \times K$ matrix $B_{it}^X = \left(x_{it,k} \sum_{s=1}^{t-1} x_{is} x_{is,k}\right) \left(x_{it,l} \sum_{s=1}^{t-1} x'_{is} x_{is,l}\right)$. The first term on the right-hand side in (2.33) has the same probability limit as $N^{-1}\widetilde{V}_{k,l}^*$, the limiting covariance matrix element $\Psi_{k,l}^*$. In contrast to LM^*, however, the variance estimator of the regression-based test involves additional quadratic forms such as $v_i' B_{it}^X v_i$, contributing to the estimator. Since, in a setup in which T is fixed,

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T} \epsilon_{it}^2 v_i' B_{it}^X v_i = O_p\left(N^{-1/2}\right),$$

the variance estimator remains consistent. In small samples, however, the additional term results in a bias of the variance estimator and may deteriorate the power of the regression-based test. See the appendix for details about the above result and the Monte Carlo experiments in Section 2.6.

We now proceed to examine the local power of the Δ statistic of PY in model (2.6) and (2.7) under the sequence of local alternatives (2.32). In our homoskedastic setup, the dispersion statistic becomes

$$\widetilde{S} = \sum_{i=1}^{N} \left(\widetilde{\beta}_i - \widetilde{\beta}\right)' \left(\frac{X_i' X_i}{\widetilde{\sigma}^2}\right) \left(\widetilde{\beta}_i - \widetilde{\beta}\right),$$

with $\widetilde{\beta}$ as the OLS estimator in (2.8) as above. Using this expression, the $\widehat{\Delta}$ statistic is computed as in (2.3). The next theorem presents the asymptotic distribution of the $\widehat{\Delta}$ statistic under the local alternatives as specified above. This result follows directly from Section 3.2 in PY.

Theorem 2.7 *Under Assumptions 2.3, 2.4, and the sequence of local alternatives (2.32)*

$$\widehat{\Delta} \xrightarrow{d} \mathcal{N}(\lambda, 1),$$

as $N \to \infty$, $T \to \infty$, provided $\sqrt{N}/T \to 0$, where $\lambda = \Lambda' c / \sqrt{2K}$ and Λ is a $K \times 1$ vector with typical element

$$\Lambda_k = \frac{1}{\sigma^2} \operatorname*{plim}_{N,T \to \infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it,k}^2,$$

for $k = 1, \ldots, K$.

In Theorem 2.7, the mean of the limiting distribution of $\widehat{\Delta}$ is slightly different from the result in Section 3.2 in PY. Here, v_i is random and independently distributed from the regressors and, therefore, the second term of the respective expression in PY is zero.

Remark 2.8 Theorems 2.5 and 2.6 imply that the local power is driven by the second moments of the regressors since

$$\Psi_{k,l} = \frac{1}{2\sigma^4} \lim_{N,T\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left(\frac{1}{T} \sum_{t=1}^{T} x_{it,k} x_{it,l} \right)^2 .$$

To illustrate the above findings, consider the simplest framework with only a single regressor. Then, according to Theorem 2.6, the non-centrality parameter becomes

$$\mu = \left(\frac{c^2}{2\sigma^4} \right) \lim_{N,T\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\left(\frac{1}{T} \sum_{t=1}^{T} x_{it}^2 \right)^2 \right] .$$

Suppose x_{it} is i.i.d. over time and independently distributed across i with uniformly bounded fourth moments. Let $\mathbb{E}\left[x_{it}\right] = 0$ and $\mathbb{E}\left[x_{it}^2\right] = \sigma_{i,x}^2$. That is, the regressor is assumed to have a unit-specific variation which is constant over time for a given unit. We then obtain

$$\mathbb{E}\left[\left(\frac{1}{T} \sum_{t=1}^{T} x_{it}^2 \right)^2 \right] = \left(\sigma_{i,x}^2 \right)^2 + O\left(T^{-1} \right),$$

implying $\mu = c^2/2\sigma^4 \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \left(\sigma_{i,x}^2 \right)^2$. To gain further insight, we think of $\left(\sigma_{i,x}^2 \right)^2$ as being randomly distributed in the cross-section such that

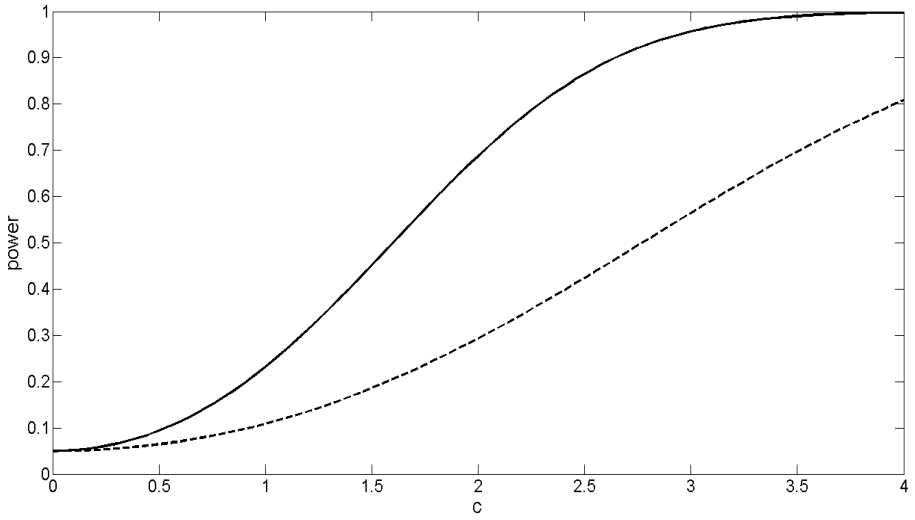
$$\mu = \frac{c^2}{2\sigma^4} \mathbb{E}\left[\left(\sigma_{i,x}^2 \right)^2 \right] = \frac{c^2}{2\sigma^4} \left(Var\left[\sigma_{i,x}^2 \right] + \left(\mathbb{E}\left[\sigma_{i,x}^2 \right] \right)^2 \right). \tag{2.34}$$

Similarly, under these assumptions, we find

$$\lambda = \frac{c}{\sigma^2 \sqrt{2}} \mathbb{E}\left[\sigma_{i,x}^2 \right]. \tag{2.35}$$

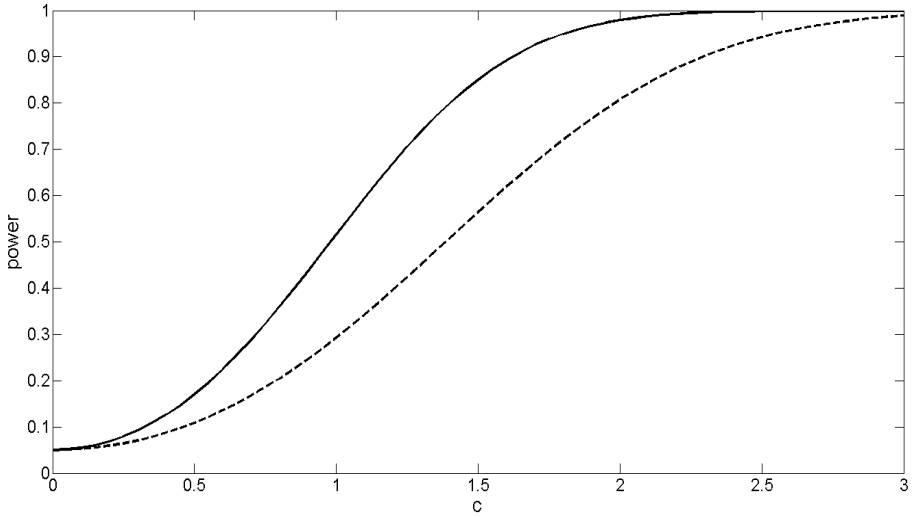
Comparing the mean of the normal distribution of the Δ statistic in (2.35) with the non-centrality parameter of the asymptotic χ_1^2 distribution of the LM statistic in (2.34), we see that the main difference between the two tests is that the variance of $\sigma_{i,x}^2$ contributes to the power of the LM statistic but not to the power of the Δ test. If $Var\left[\sigma_{i,x}^2 \right] = 0$ such that $\sigma_{i,x}^2 = \sigma_x^2$ for all i, the LM test and the Δ test have the same asymptotic power in this example. If, however, $Var\left[\sigma_{i,x}^2 \right] > 0$, so that there is variation in the variance of the regressor in the cross-section, the LM test has larger asymptotic power. To illustrate this point, we examine the local asymptotic power functions of the LM and the Δ test for two cases, using the expressions in (2.34) and (2.35). Figure 2.1 shows the local asymptotic power of the LM (solid line) and the Δ test (da-

Figure 2.1: Asymptotic local power of the LM and the Δ test when $\sigma_{i,x}^2 \sim \chi_1^2$



Note: Solid line: local power of LM test according to (2.34). Dashed line: local power of Δ test according to (2.35). Here, $\sigma^2 = 1$ and the variances $\sigma_{i,x}^2$ are distributed independently across i.

Figure 2.2: Asymptotic local power of the LM and the Δ test when $\sigma_{i,x}^2 \sim \chi_2^2$.



Note: Solid line: local power of LM test implied by (2.34). Dashed line: local power of Δ test implied by (2.35). Here, $\sigma^2 = 1$ and the variances $\sigma_{i,x}^2$ are distributed independently across i.

shed line) as a function of c when $\sigma_{i,x}^2$ has a χ_1^2 distribution. Figure 2.2 repeats this exercise for $\sigma_{i,x}^2$ drawn from a χ_2^2 distribution. In both cases, the LM test has larger asymptotic power. The power gain is substantial for the first case, but diminishes for the second. This pattern is expected, as the variance of $\sigma_{i,x}^2$ contributes relatively more to the non-centrality parameter in the first specification.

This discussion exemplifies the difference between the LM-type tests and the Δ statistic in terms of the local asymptotic power in a simplified framework. The analysis suggests that the LM-type tests are particularly powerful in an empirically relevant setting in which there is non-negligible variation in the variances of the regressors between panel units.
Having studied the large samples properties of the LM tests under the null and the alternative hypothesis in our model, we now evaluate the finite-sample size and power properties of the LM-type tests in a Monte Carlo experiment.

2.6 Monte Carlo experiments

2.6.1 Design

After deriving LM-type tests in the random coefficient model, we now turn to study the small-sample properties of the proposed test and its variants. The aim of this section is to evaluate the performance of the tests in terms of their empirical size and power in several different setups, relating to the theoretical discussion of Sections 2.3 - 2.5. We consider the following test statistics: the original LM statistic presented in Theorem 2.1, the adjusted LM statistic that adjusts the information matrix to account for fourth moments of the error distribution (see Corollary 2.1), the score-modified LM statistic (see Theorem 2.3) and the regression-based, heteroskedasticity-robust LM statistic (see Theorem 2.4). As a benchmark, we consider PY's statistic $\widetilde{\Delta}_{\text{adj}}$ given in (2.4). Following the notes in Table 1 in PY, the test using $\widetilde{\Delta}_{\text{adj}}$ is carried out as a two-sided test. In addition, the CLM test in (2.19) is included, which is also a two-sided test.
We consider the following data-generating process with normally distributed errors as the stan-

dard design:

$$
\begin{aligned}
y_{it} &= \alpha_i + x_{it}'\beta_i + \epsilon_{it}, && (2.36)\\
\epsilon_{it} &\overset{iid}{\sim} \mathcal{N}(0,1), && (2.37)\\
\alpha_i &\overset{iid}{\sim} \mathcal{N}(0,0.25),\\
x_{it,k} &= \alpha_i + \vartheta^x_{it,k},\ k=1,2,3,\\
\vartheta^x_{it,k} &\overset{iid}{\sim} \mathcal{N}\left(0,\sigma^2_{ix,k}\right),\\
\beta_i &\overset{iid}{\sim} \mathcal{N}_3(\iota_3,\Sigma_v),\\
\Sigma_v &= \begin{bmatrix} 0.03 & 0 & 0 \\ 0 & 0.02 & 0 \\ 0 & 0 & 0.01 \end{bmatrix}, && (2.38)
\end{aligned}
$$

where $i = 1,2,\ldots,N$, $t = 1,2,\ldots,T$. As discussed in Section 2.5 the variances of the regressors play an important role. In our benchmark specification we generate the variances as

$$
\begin{aligned}
\sigma^2_{ix,k} &= 0.25 + \eta_{i,k}\\
\eta_{i,k} &\overset{iid}{\sim} \chi^2_1, && (2.39)
\end{aligned}
$$

The choice of the χ^2 distribution for $\sigma^2_{ix,k}$ is made analogous to the Monte Carlo experiment in PY, see page 63 in PY. We then consider variations of this specification below. All results are based on 5,000 Monte Carlo replications. We choose

$$
\begin{aligned}
N &\in \{10,20,30,50,100,200\},\\
T &\in \{10,20,30\},
\end{aligned}
$$

as we would like to study the small-sample properties of the test procedures when the time dimension is small.

In our first set of Monte Carlo experiments the errors are normally distributed; therefore we focus on the standard LM test. We also include their respective heteroskedasticity-robust regression variants for this exercise.

2.6.2 Results: normally distributed errors

Panel A of Table 2.1 (see Appendix 2.B) shows the rejection frequencies when the null hypothesis is true. The $\widetilde{\Delta}_{\text{adj}}$ test has rejection frequencies close to the nominal size of 5% for

all combinations of N and T, while the CLM test rejects the null hypothesis too often, in particular for small N. Deviations from the nominal size for the the standard LM test and the regression-based test are small and disappear as N increases, as expected from Theorem 2.1.

Panel B of Table 2.1 shows the corresponding rejections frequencies under the alternative hypothesis. The LM test outperforms the $\widetilde{\Delta}_{\text{adj}}$ and the CLM test in general. This observation holds in particular for $T = 10$ where the power gain is considerable. The LM-OPG variant, although as powerful as the $\widetilde{\Delta}_{\text{adj}}$ test for $T = 10$, suffers from a power loss relative to the standard LM test. This power loss may be due to the small-sample bias of the variance estimator, see Remark 2.7.

Following Remark 2.5 the variants of the LM tests according to Theorems 3 and 4 are computed as follows. First, the individual-specifc fixed effects α_i are eliminated by transforming the data using orthogonal deviations (see Arellano and Bover (1995)). The LM statistics are then computed using the transformed data. The results for the within transformation (see Panel A of Table 2.2) and the forward orthogonalization (not reported) indicate that the LM test is sensitive to the implied serial correlation in the error term when applying the usual within-group estimator. As illustrated in section 2.5, the power gain of the LM test is directly related to the variation in the variances of the regressors. We therefore change the above design with respect to (2.39) as follows:

$$\begin{aligned} \sigma^2_{ix,k} &= 0.25 + \eta_{i,k}, \\ \eta_{i,k} &\overset{iid}{\sim} \chi^2_2, \end{aligned} \tag{2.40}$$

such that $\eta_{i,k}$ has now a χ^2 distribution with 2 degrees of freedom. As the empirical size is very similar to the previous setup, we focus on the rejection rates under the alternative presented in panel B of Table 2.2. By comparing panel B of Table 2.1 and the rejection rates in panel B of Table 2.2, this exercise illustrates the analysis underlying Figures 2.1 and 2.2 in small samples. We see that the power gain of the LM test is still sizeable for $T = 10$. As T increases, however, the gap between the LM test and the $\widetilde{\Delta}_{\text{adj}}$ test in term of their empirical power becomes smaller.

2.6.3 Results: non-normal errors

We now investigate the LM test when the errors are no longer normally distributed, thereby building on the results of Section 2.4.1. The errors in (2.37) are generated from a t-distribution with 5 degrees of freedom, scaled to have unit variance. All other specifications of the standard

design remain unchanged. In addition to the statistics already considered, we now include the adjusted LM statistic (see corollary 2.1) and the score-modified statistic (see Theorem 2.3).
Panel A in Table 2.3 (see Appendix 2.B) reports the rejection frequencies under the null hypothesis in this case. We notice that the LM test has substantial size distortions when T is fixed and N increases, which is expected from Theorem 2.2. However, the adjusted LM statistic LM_{adj} and the modified score statistic LM^* are both successful in controlling the type-I error.
Panel B of Table 2 shows rejection frequencies under the alternative hypothesis. The power gain of the LM test relative to the $\widetilde{\Delta}_{\text{adj}}$ test is noticeable when $T = 10$ or $T = 20$. Qualitatively, we draw the same conclusions when the errors are χ^2 with two degrees of freedom, centered and standardized to have mean zero and variance equal to one. These results are presented in Table 2.4. Here, the size distortions of the LM test for small values of the time dimension are more pronounced, and the adjusted versions are again able to provide reliable inference.

2.6.4 Additional simulation results

When we first introduced the LM test in Section 2.3, we made the simplifying assumption that the random components of the coefficients have a diagonal covariance matrix. To study the properties of the test for correlated random coefficients we now allow for non-zero off-diagonal elements in the matrix Σ_v given in (2.38) above. The covariances are chosen such that the first and second component of v_i have correlation equal to 0.5, the second and third component have correlation 0.25 and the first and third have correlation equal to zero. The variances are as above. All other parameters are chosen as in the standard design. Rejection frequencies under this specification of the alternative are presented in Table 2.5. Since the results are very similar to the previous results when Σ_v is diagonal, the LM test remains powerful under such alternatives.

2.7 Concluding remarks

In this chapter we examine the problem of testing slope homogeneity in a panel data model. We develop testing procedures using the LM principle. Several variants are considered that robustify the original LM test with respect to non-normality and heteroscedasticity. By studying the local power we identify cases where the LM-type tests are particularly powerful relative to existing tests. In sum, our Monte Carlo experiments suggest that the LM test are powerful testing procedures to detect slope homogeneity in short panels in which the time dimension is

small relative to the cross-section dimension. The LM approach suggested in this chapter may be extended in future research by allowing for dynamic specifications with lagged dependent variables and cross sectionally or serially correlated errors.

Appendix to Chapter 2

2.A Proofs

To economize notation we use $\sum_i$ and $\sum_t$ instead of full expressions $\sum_{i=1}^{N}$ and $\sum_{t=1}^{T}$ throughout this appendix. Moreover, tr $[A]$ denotes the trace of the square matrix A.

Preliminary results

We first present an important result concerning the asymptotic effect of the estimation error $\widetilde{\beta} - \beta$ on the test statistics. Define

$$A_i^{(k)} = X_i^{(k)} X_i^{(k)\prime} - \left(\frac{1}{NT} \sum_i X_i^{(k)\prime} X_i^{(k)} \right) I_T.$$

Lemma 2.2 *Let* $R_{XAX}^{(k)} = \sum_i X_i' A_i^{(k)} X_i$ *and* $R_{XAu}^{(k)} = \sum_i X_i' A_i^{(k)} u_i$ *for* $k = 1, ..., K$. *Furthermore let*

$$R_N^{(k)} = \left(\frac{\widetilde{\sigma}^4}{\sigma^4} \right) \frac{1}{2\sigma^2} \left(\left(\widetilde{\beta} - \beta \right)' R_{XAX}^{(k)} \left(\widetilde{\beta} - \beta \right) - 2 \left(\widetilde{\beta} - \beta \right)' R_{XAu}^{(k)} \right),$$

for $k = 1, ..., K$. *Under Assumptions 2.1, 2.2 and the null hypothesis the following properties hold if* T *is fixed:*

(i) $R_{XAX}^{(k)} = O_p(N)$,

(ii) $R_{XAu}^{(k)} = O_p\left(N^{1/2}\right)$,

(iii) $R_N^{(k)} = O_p(1)$,

for $k = 1, ..., K$.

Proof. (i) Using the definition of $A_i^{(j)}$ yields

$$R_{XAX}^{(k)} = \sum_i X_i' \left(X_i^{(k)} X_i^{(k)\prime} \right) X_i - \frac{1}{NT} \left(\sum_i \sum_t x_{it,k}^2 \right) \left(\sum_i X_i' X_i \right).$$

The first term is a $K \times K$ matrix with typical (l, m) element

$$\sum_i \left(\sum_t x_{it,l} x_{it,k} \right) \left(\sum_t x_{it,m} x_{it,k} \right) = O_p(N),$$

as a consequence of Assumption 2.2, while $\sum_i \sum_t x_{it,k}^2 / NT = O_p(1)$ and $\sum_i X_i' X_i = O_p(N)$.
(ii) Recall that under the null hypothesis, $u_i = \epsilon_i$. Thus

$$R_{XAu}^{(k)} = \sum_i \left(X_i' X_i^{(k)} \right) \left(X_i^{(k)\prime} u_i \right) - \frac{1}{NT} \left(\sum_i \sum_t x_{it,k}^2 \right) \left(\sum_i X_i' u_i \right).$$

The first and the second term are $O_p\left(N^{1/2}\right)$ by a the central limit theorem (CLT) for independent random variables and Assumption 2.2.
(iii) Combining **(i)** and **(ii)** together with the fact that $\sqrt{N}\left(\widetilde{\beta}-\beta\right)=O_p(1)$ yields the result.
■

Lemma 2.3 *Under Assumptions 2.3, 2.4 and the null hypothesis the following properties hold for $N\to\infty$ and $T\to\infty$:*

(i) $R_{XAX}^{(k)}=O_p\left(NT^2\right)$,

(ii) $R_{XAu}^{(k)}=O_p\left(N^{1/2}T^{3/2}\right)$,

(iii) $R_{NT}^{(k)}=O_p\left(T\right)$, *which is defined as $R_N^{(k)}$ in Lemma 2.2,*

for $k=1,...,K$.

Proof. Following the proof of Lemma 2.2 the element of the first term of $R_{XAX}^{(k)}$ is $O_p\left(NT^2\right)$, whereas the second term is $O_p\left(NT\right)$ by Assumption 2.4 which proves statement **(i)**. Notice in **(ii)** $R_{XAu}^{(k)}$ has two terms as in Lemma 2.2, where the first one has zero mean and variance of order T^3. Therefore by Lemma 1 in Baltagi, Feng, and Kao (2011) we have that $X_i'X_i^{(j)}X_i^{(j)\prime}u_i=O_p(T^{3/2})$ and by Lemma 2 in PY that $\sum_i\left(X_i'X_i^{(j)}\right)\left(X_i^{(j)\prime}u_i\right)=O_p\left(N^{1/2}T^{3/2}\right)$ and $\left(\sum_i X_i'u_i\right)$ $=O_p\left(N^{1/2}T^{1/2}\right)$. Using these results and the fact that $\sqrt{NT}\left(\widetilde{\beta}-\beta\right)=O_p\left(1\right)$ implies **(iii)**.
■

Proofs of the main results

Proof of Lemma 2.1

We use the following rules for matrix differentiations:

$$\frac{\partial\ell}{\partial\theta_k}=-\frac{1}{2}\text{tr}\left[\Omega^{-1}\frac{\partial\Omega}{\partial\theta_k}\right]+\frac{1}{2}\left[u'\Omega^{-1}\frac{\partial\Omega}{\partial\theta_k}\Omega^{-1}u\right], \tag{2.41}$$

$$-\mathbb{E}\left[\frac{\partial\ell}{\partial\theta_k\partial\theta_l}\right]=\frac{1}{2}\text{tr}\left[\Omega^{-1}\left(\frac{\partial\Omega}{\partial\theta_k}\right)\Omega^{-1}\left(\frac{\partial\Omega}{\partial\theta_l}\right)\right], \tag{2.42}$$

for $k,l=1,2,\ldots,K+1$, see, e.g., Harville (1977) and Wand (2002).
First,

$$X_i\Sigma_vX_i'=\sum_k\sigma_{v,k}^2X_i^{(k)}X_i^{(k)\prime},$$

with $X_i^{(k)}$ denoting the k-th column vector of X_i. Hence

$$\begin{bmatrix} X_1\Sigma_vX_1' & & 0 \\ & \ddots & \\ 0 & & X_N\Sigma_vX_N' \end{bmatrix}=\sum_k\sigma_{v,k}^2A_k,$$

with the $NT \times NT$ matrix ,

$$A_k = \begin{bmatrix} X_1^{(k)} X_1^{(k)\prime} & & 0 \\ & \ddots & \\ 0 & & X_N^{(k)} X_N^{(k)\prime} \end{bmatrix},$$

for $k = 1, \ldots, K$, and $X_i^{(k)}$ denotes the k-th column of the $T \times K$ matrix X_i. Thus,

$$\Omega = \sum_k \sigma_{v,k}^2 A_k + \sigma^2 I_{NT},$$

and

$$\frac{\partial \Omega}{\partial \theta_k} = \begin{cases} A_k, & \text{for } k = 1, 2, \ldots, K, \\ I_{NT}, & \text{for } k = K+1. \end{cases}$$

Under the null hypothesis we have $\Omega = \sigma^2 I_{NT}$. Using (2.41) we obtain

$$\left.\frac{\partial \ell}{\partial \theta_k}\right|_{H_0} = \begin{cases} -\frac{1}{2\widetilde{\sigma}^2} \text{tr}\left[A_k\right] + \frac{1}{2\widetilde{\sigma}^4} \widetilde{u}' A_k \widetilde{u}, & \text{for } k = 1, 2, \ldots, K, \\ 0, & \text{for } k = K+1. \end{cases}$$

where

$$\begin{aligned} \widetilde{\sigma}^2 &= \frac{1}{NT} \widetilde{u}' \widetilde{u}, \\ \widetilde{u} &= \left(I_{NT} - X \left(X'X\right)^{-1} X'\right) y. \end{aligned}$$

The representation of the score vector follows from

$$\text{tr}\left[A_k\right] = \sum_i \sum_t X_{it,k}^2 = X^{(k)\prime} X^{(k)},$$

where $X^{(k)}$ denotes the k-th column of the $NT \times K$ matrix X.
Similarly, (2.42) yields

$$-\mathbb{E}\left[\frac{\partial \ell}{\partial \theta_k \partial \theta_l}\right]\Bigg|_{H_0} = \begin{cases} \frac{1}{2\sigma^4} \text{tr}\left[A_k A_l\right], & \text{for } k, l = 1, 2, \ldots, K, \\ \frac{1}{2\sigma^4} X^{(k)\prime} X^{(k)}, & \text{for } k = 1, 2, \ldots, K, \text{ and } l = K+1, \\ \frac{NT}{2\sigma^4}, & \text{for } k = l = K+1. \end{cases}$$

Using the fact that A_k and A_l are block-diagonal,

$$\text{tr}\left[A_k A_l\right] = \sum_i \text{tr}\left[\left(X_i^{(k)} X_i^{(k)\prime}\right)\left(X_i^{(l)} X_i^{(l)\prime}\right)\right] = \sum_i \left(X_i^{(k)\prime} X_i^{(l)}\right)^2,$$

where $X_i^{(k)}$ denotes the i-th column of X_i, which yields the form of the information matrix presented in the lemma.

Proof of Theorem 2.1

Recall that

$$A_i^{(k)} = X_i^{(k)} X_i^{(k)\prime} - \left(\frac{1}{NT} \sum_i X_i^{(k)\prime} X_i^{(k)} \right) I_T,$$

and rewrite the elements of the scores as

$$\widetilde{s}_k = \left(\frac{\widetilde{\sigma}^4}{\sigma^4} \right) \frac{1}{2\sigma^4} \sum_i \widetilde{u}_i' A_i^{(k)} \widetilde{u}_i,$$

for $k = 1, ..., K$. Since $\widetilde{u}_i = u_i - X_i(\widetilde{\beta} - \beta)$ we have

$$\frac{1}{\sqrt{N}} \widetilde{s}_k = \frac{1}{\sqrt{N}} \left(\frac{\sigma^4}{\widetilde{\sigma}^4} \right) \frac{1}{2\sigma^4} \sum_i u_i' A_i^{(k)} u_i + \frac{1}{\sqrt{N}} R_N^{(k)},$$

where $R_N^{(k)} = O_p(1)$ from Lemma 2.2. Since $\sum_i \text{tr}\left[A_i^{(k)} \right] = 0$ it follows that $\mathbb{E}(u_i' A_i^{(k)} u_i) = 0$ and, therefore,

$$\lim_{N \to \infty} \mathbb{E}\left(\frac{1}{\sqrt{N}} \widetilde{s} \right) = 0.$$

The covariances are obtained as

$$\begin{aligned}
&Cov\left(u_i' A_i^{(k)} u_i, u_i' A_i^{(l)} u_i \mid X \right) = 2\sigma^4 \text{tr}\left[A_i^{(k)} A_i^{(l)} \right] \\
&= 2\sigma^4 \left(X_i^{(k)\prime} X_i^{(l)} \right)^2 - \left(\frac{1}{NT} \sum_i X_i^{(k)\prime} X_i^{(k)} \right) \left(X_i^{(l)\prime} X_i^{(l)} \right) \\
&- \left(\frac{1}{NT} \sum_i X_i^{(l)\prime} X_i^{(l)} \right) \left(X_i^{(k)\prime} X_i^{(k)} \right) + T \left(\frac{1}{NT} \sum_i X_i^{(k)\prime} X_i^{(k)} \right) \left(\frac{1}{NT} \sum_i X_i^{(l)\prime} X_i^{(l)} \right),
\end{aligned}$$

and since $u_i' A_i^{(k)} u_i$ is independent of $u_j' A_i^{(l)} u_j$ for all $i \neq j$ conditional on X,

$$\begin{aligned}
&\left(\frac{1}{2\sigma^4} \right)^2 Cov\left(\sum_i u_i' A_i^{(k)} u_i, \sum_i u_i' A_i^{(l)} u_i \;\middle|\; X \right) \\
&= \frac{1}{2\sigma^4} \left(\sum_i \left(X_i^{(k)\prime} X_i^{(l)} \right)^2 - \frac{1}{NT} \left(\sum_i X^{(k)\prime} X^{(k)} \right) \left(\sum_i X_i^{(l)\prime} X_i^{(l)} \right) \right) \\
&= V_{k,l}.
\end{aligned}$$

The Liapounov condition in the central limit theorem for independent random variables (see White (2001), Theorem 5.10) is satisfied by Assumption 2 and therefore

$$\left(\frac{1}{N} \widetilde{V} \right)^{-1/2} \left(\frac{1}{\sqrt{N}} \widetilde{s} \right) \xrightarrow{d} \mathcal{N}(0, I_K),$$

where $\widetilde{V}$ replaces σ^4 in V by $\widetilde{\sigma}^4$. By the formula for the partitioned inverse

$$\left\{ \mathcal{I}(\widetilde{\sigma}^2)^{-1} \right\}_{1:K,1:K} = \widetilde{V}^{-1},$$

where $\{\cdot\}_{1:K,1:K}$ denotes the upper-left $K \times K$ block of the matrix, it follows finally that

$$\widetilde{\mathcal{S}}'\mathcal{I}(\widetilde{\sigma}^2)^{-1}\widetilde{\mathcal{S}} = \widetilde{s}'\widetilde{V}^{-1}\widetilde{s} \xrightarrow{d} \chi^2_K.$$

Proof of Theorem 2.2

The proof proceeds in three steps: **(i)** we derive the variance-covariance matrix of the score vector, **(ii)** we establish the asymptotic normality of the score and **(iii)** we use these results to establish the asymptotic distribution of the LM statistic.
(i) Define the $K \times 1$ vector s with typical element

$$s_k = u_i' A_i^{(k)} u_i. \tag{2.43}$$

Using standard results for quadratic forms (see for example Ullah (2004), appendix A.5),

$$\mathbb{E}\left[u_i' A_i^{(k)} u_i \middle| X\right] = \sigma^2 \text{tr}\left[A_i^{(k)}\right]$$

$$\begin{aligned}\mathbb{E}\left[\left(u_i' A_i^{(k)} u_i\right)\left(u_i' A_i^{(l)} u_i\right)\middle| X\right] = {} & 2\sigma^4 \text{tr}\left[A_i^{(k)} A_i^{(l)}\right] + \sigma^4 \text{tr}\left[A_i^{(k)}\right] \text{tr}\left[A_i^{(l)}\right] \\ & + \left(\mu_u^{(4)} - 3\sigma^4\right) a_i^{(k)\prime} a_i^{(l)},\end{aligned}$$

where $a_i^{(k)}$ is a vector consisting of the main diagonal elements of the matrix $A_i^{(k)}$ and $\mu_u^{(4)}$ denotes the fourth moment of u_{it}. Since

$$\mathbb{E}\left[u_i' A_i^{(k)} u_i \middle| X\right] \mathbb{E}\left[u_i' A_i^{(l)} u_i \middle| X\right] = \sigma^4 \text{tr}\left[A_i^{(k)}\right] \text{tr}\left[A_i^{(l)}\right],$$

we have

$$Cov\left(u_i' A_i^{(k)} u_i, u_i' A_i^{(l)} u_i \middle| X\right) = 2\sigma^4 \text{tr}\left[A_i^{(k)} A_i^{(l)}\right] + \left(\mu_u^{(4)} - 3\sigma^4\right) a_i^{(k)'} a_i^{(l)}.$$

Due to the independence of $u_i' A_i^{(k)} u_i$ and $u_j' A_j^{(l)} u_j$ for $i \neq j$, it follows that

$$Cov\left(\sum_i u_i' A_i^{(k)} u_i, \sum_i u_i' A_i^{(l)} u_i \middle| X\right) = 2\sigma^4 \sum_i \text{tr}\left[A_i^{(k)} A_i^{(l)}\right] + \left(\mu_u^{(4)} - 3\sigma^4\right) \sum_i a_i^{(k)\prime} a_i^{(l)}.$$

Inserting the expression for $tr\left[A_i^{(k)} A_i^{(l)}\right]$, we determine the (k,l) element of the covariance matrix of s as

$$\begin{aligned} V^s_{k,l} &= \frac{1}{2\sigma^4}\left(\sum_i\left(\sum_t x_{it,j}x_{it,l}\right)^2 - \frac{1}{NT}\left(\sum_i\sum_t x^2_{it,k}\right)\left(\sum_i\sum_t x^2_{it,k}\right)\right) \\ &\quad + \left(\frac{\mu_u^{(4)} - 3\sigma^4}{(2\sigma^4)^2}\right)\sum_i\sum_t\left(x^2_{it,k} - \frac{1}{NT}\sum_i\sum_t x^2_{it,k}\right)\left(x^2_{it,k} - \frac{1}{NT}\sum_i\sum_t x^2_{it,l}\right) \\ &= V^s_{1,k,l} + V^s_{2,k,l}. \end{aligned} \tag{2.44}$$

(ii) To verify that a central limit theorem applies to s, let $\lambda \in \mathbb{R}^k$, $||\lambda|| = 1$. Following Jiang

(1996), note that

$$\begin{aligned} u_i A_i^{(k)} u_i - \sigma^2 \text{tr}\left[A_i^{(k)}\right] &= \sum_t a_{i,tt}^{(k)} \left(u_{it}^2 - \sigma^2\right) + \sum_t \sum_{s \neq t} a_{i,ts}^{(k)} u_{is} u_{it} \\ &= \sum_t a_{i,tt}^{(k)} \left(u_{it}^2 - \sigma^2\right) + 2 \sum_t \left(\sum_{s=1}^{t-1} a_{i,ts}^{(k)} u_{is} \right) u_{it}, \end{aligned}$$

and

$$\frac{1}{2\sigma^4} \sum_i \left(u_i A_i^{(k)} u_i - \text{tr}\left[A_i^{(k)}\right] \right) = \frac{1}{2\sigma^4} \sum_i \sum_t Z_{i,t}^{(k)},$$

where

$$Z_{i,t}^{(k)} \equiv a_{i,tt}^{(k)} \left(u_{it}^2 - \sigma^2\right) + 2 \left(\sum_{s=1}^{t-1} a_{i,ts}^{(k)} u_{is} \right) u_{it},$$

with $a_{i,tt}^{(k)} = x_{it,j}^2 - \overline{x}_{NT}^{(k)}$ and $\overline{x}_{NT}^{(k)} = \frac{1}{NT} \sum_i \sum_t x_{it,k}^2$. Let $\mathcal{F}_{i,t}$ be the sigma-field generated by $(u_{i1}, \ldots, u_{it})$. Then $\left(Z_{i,t}^{(j)}, \mathcal{F}_{i,t}\right)$ is a martingale difference sequence for $k = 1, \ldots, K$. Using the Cramer-Wold device for the properly normalized elements of the score vector,

$$\frac{1}{T} \lambda' s = \sum_k \frac{\lambda_k}{2\sigma^4} \left(\sum_i \frac{1}{T} \sum_t Z_{i,t}^{(k)} \right) = \sum_i \xi_{i,N},$$

with $\xi_{i,N} = \frac{1}{T} \sum_t \sum_k \left(\lambda_k / 2\sigma^2\right) Z_{i,t}^{(k)} = \frac{1}{2\sigma^4} \frac{1}{T} \lambda' Z_{i,N}$. Note that $(\xi_{i,N}, \mathcal{F}_{i,N})$ is an md array where $\mathcal{F}_{i,N} = \sigma(\xi_{i-1,t}, \xi_{i-2,t} ..., \xi_{i-1,t-1}, ..)$ for all $t = 1, ..., T$ and

$$E\left[\xi_{i,N}^2\right] = \frac{1}{4\sigma^8} E\left[Z_{i,N}' \lambda \lambda' Z_{i,N}\right] = \frac{1}{4\sigma^8 T^2} \lambda' V_i^s \lambda. \tag{2.45}$$

Then the CLT for md arrays (Davidson (1994), Thm. 24.3) applies to the normalized sequence

$$\zeta_{i,N} := \frac{\frac{1}{\sqrt{N}} 2\sigma^4 \xi_{i,N}}{\sqrt{\mathcal{V}_{N,T}}}, \tag{2.46}$$

where $\mathcal{V}_{N,T} = \frac{1}{NT^2} \lambda' V^s \lambda$, and $V^s = \sum_i V_i^s$, if two conditions are satisfied:

$$\sum_i \zeta_{i,N}^2 \xrightarrow{p} 1,$$

$$\max_{1 \leq i \leq N} |\zeta_{i,N}| \xrightarrow{p} 0.$$

Regarding the first condition, from Lemma 1 in Hansen (2007)

$$\frac{1}{NT^2} \sum_i \left(\xi_{i,N}\right)^2 \xrightarrow{p} \frac{1}{4\sigma^8} \lim_{N,T \to \infty} \frac{1}{NT^2} \lambda' V^s \lambda = \frac{1}{4\sigma^8} \lim_{N,T \to \infty} \mathcal{V}_{N,T}, \tag{2.47}$$

provided that $E\left[\left|\xi_{i,N}\right|^3\right] < C < \infty$ for all i, N and T. To study wether $E\left[\left|\xi_{i,N}\right|^3\right]$ is uniformly

bounded it suffices to consider $\xi_{i,N}$ elementwise, i.e.,

$$\begin{aligned}
&\left(\frac{1}{2\sigma^4}\right)^3 \mathbb{E}\left[\frac{1}{T^3}\left|\sum_t Z_{i,t}^{(k)}\right|^3\right] \\
= &\left(\frac{1}{2\sigma^4}\right)^3 \mathbb{E}\left[\frac{1}{T^3}\left|\sum_t \left(a_{i,tt}^{(k)}\left(u_{it}^2-\sigma^2\right)+2\left(\sum_{s=1}^{t-1} a_{i,ts}^{(k)} u_{is}\right)u_{it}\right)\right|^3\right].
\end{aligned}$$

Using Hölders inequality we also have

$$\left(\frac{1}{2\sigma^4}\right)^3 \mathbb{E}\left[\frac{1}{T^3}\left|\sum_t Z_{i,t}^{(k)}\right|^3\right] \leq \left(\frac{1}{2\sigma^4}\right)^3 \left(\sum_t \frac{1}{T}\left\|Z_{i,t}^{(k)}\right\|_3\right)^3.$$

Making use of independence of the u_{it} and the triangle inequality

$$\frac{1}{T}\left\|Z_{i,t}^{(k)}\right\|_3 \leq \frac{1}{T}\left\|a_{i,tt}^{(k)}\right\|_3 \left\|\left(u_{it}^2-\sigma^2\right)\right\|_3 + \frac{1}{T}\left\|2\left(\sum_{s=1}^{t-1} a_{i,ts}^{(k)} u_{is}\right)\right\|_3 \left\|u_{it}\right\|_3.$$

With $\mathbb{E}\left(u_{it}^6\right)$ being finite and uniformly bounded, $\left\|(u_{it}^2-\sigma^2)\right\|_3$ and $\left\|u_{it}\right\|_3$ are uniformly bounded in i and t. From Assumption 2.4 we have that $\left\|a_{i,tt}^{(j)}\right\|_3$ is uniformly bounded. Finally, notice that

$$\frac{1}{T}\left\|\left(\sum_{s=1}^{t-1} a_{i,ts}^{(k)} u_{is}\right)\right\|_3 \leq \frac{1}{T}\sum_{s=1}^{t-1}\left\|a_{i,tt}^{(k)}\right\|_3 \left\|u_{it}\right\|_3,$$

is also uniformly bounded by the same argument. Putting these results together $E\left[\left|\xi_{i,N}\right|^3\right]$ is uniformly bounded as well.
Therefore, by (2.47)

$$\sum_i \zeta_{i,N}^2 \xrightarrow{p} \operatorname*{plim}_{N,T\to\infty}\left(\frac{2\sigma^4}{\sqrt{\lambda' V^s \lambda}}\right)^2 \xi_{i,N}^2 = 1.$$

Regarding the second condition notice that since $\mathcal{V}_{N,T}$ is uniformly bounded it is sufficient to prove that $\max_{1\leq i\leq N}\left|\frac{1}{\sqrt{N}}\xi_{i,N}\right| \xrightarrow{p} 0$. Now for any $\varepsilon>0$,

$$\begin{aligned}
\Pr\left(\frac{1}{\sqrt{N}}\max_{1\leq i\leq N}\left|\xi_{i,N}\right|>\varepsilon\right) &\leq \sum_i \Pr\left(\frac{1}{\sqrt{N}}\left|\xi_{i,N}\right|>\varepsilon\right) \\
&\leq \frac{1}{\varepsilon^3 N^{3/2}}\sum_i \mathbb{E}\left[\left|\xi_{i,N}\right|^3\right] = O\left(N^{-1/2}\right),
\end{aligned}$$

where the first inequality follows from Bonferroni's inequality, the second uses the generalized Markov inequality and the last equality is due to the uniform boundnes of $\mathbb{E}\left[\left|\xi_{i,N}\right|^3\right]$, as shown above. This completes the proof of **(ii)**.
(iii) Rewrite the first K elements of the score as

$$\widetilde{s} = \left(\frac{\widetilde{\sigma}^4}{\sigma^4}\right) s + R_{NT},$$

where R_{NT} is given in Lemma 2.3 and s has typical element as defined in (2.43).
By **(ii)**,

$$s' \left(V^s\right)^{-1} s \xrightarrow{d} \chi^2_K, \tag{2.48}$$

as $N \to \infty$, $T \to \infty$, where V^s has (k,l) element $V^s_{k,l}$ as in (2.44).
Under Assumptions 2.3 and 2.4

$$\begin{aligned} V_1^s &= O_p\left(NT^2\right), \\ V_2^s &= O_p\left(NT\right), \end{aligned}$$

where V_1^s and V_2^s are specified elementwise in (2.44). Given the expression for $\widetilde{V}$ in Theorem 1,

$$\frac{\widetilde{V}}{NT^2} - \frac{V_1^s}{NT^2} \xrightarrow{p} 0$$

and hence

$$s'\widetilde{V}^{-1}s - s' \left(V^s\right)^{-1} s \xrightarrow{p} 0 \tag{2.49}$$

as $N \to \infty, T \to \infty$.
The LM statistic can be expanded as

$$\begin{aligned} \text{LM} &= \widetilde{s}'\widetilde{V}^{-1}\widetilde{s} \\ &= \left(\left(\frac{\widetilde{\sigma}^4}{\sigma^4}\right) s + R_{NT}\right)' \widetilde{V}^{-1} \left(\left(\frac{\widetilde{\sigma}^4}{\sigma^4}\right) s + R_{NT}\right) \\ &= \left(\frac{\widetilde{\sigma}^4}{\sigma^4}\right) \left(s'\widetilde{V}^{-1}s\right) + O_p\left(N^{-1/2}\right). \end{aligned} \tag{2.50}$$

where the last line follows from Lemma 2.3. The theorem follows by combining (2.48), (2.49) and (2.50).

Proof of Corollary 2.1

The result follows immediately from the proof of Theorem 2.2 and the fact that

$$\widetilde{\mu}_u^{(4)} = (NT)^{-1} \sum_i \sum_t \widetilde{u}_{it}^4$$

is a consistent estimator of $\mu_u^{(4)}$.

Proof of Theorem 2.3

Using similar arguments as in the proof of Theorem 2.1,

$$\frac{1}{\sqrt{N}}\widetilde{s}^* = \frac{1}{\sqrt{N}} \left(\frac{\sigma^4}{\widetilde{\sigma}^4}\right) \begin{bmatrix} \frac{1}{\sigma^4} \sum_i \sum_t \sum_{s=1}^{t-1} x_{it,1} u_{it} x_{is,1} u_{is} \\ \vdots \\ \frac{1}{\sigma^4} \sum_i \sum_t \sum_{s=1}^{t-1} x_{it,K} u_{it} x_{is,K} u_{is} \end{bmatrix} + o_p(1). \tag{2.51}$$

Let $u^*_{it} = u_{it}/\sigma$ and $z^*_{itj} = x_{it,j} u^*_{it}$. Clearly, $\mathbb{E}\left[\sum_t \sum_{s=1}^{t-1} z^*_{itk} z^*_{isk}\right] = 0$. Since conditional on X, $\sum_t \sum_{s=1}^{t-1} z^*_{it,k} z^*_{is,k}$ and $\sum_t \sum_{s=1}^{t-1} z^*_{jtl} z^*_{jsl}$ are independent for $i \neq j$, the covariances for two

elements k and l of the vector (2.51) are

$$\begin{aligned}\mathbb{E}\left[s_k^* s_l^* \,\middle|\, X\right] &= \frac{1}{\sigma^4}\sum_i \mathbb{E}\left[\left(\sum_{t=2}^{T}\sum_{s=1}^{t-1} z_{itj}^* z_{isj}^*\right)\left(\sum_{t=2}^{T}\sum_{s=1}^{t-1} z_{itl}^* z_{isl}^*\right)\middle| X\right] \\ &= \frac{1}{\sigma^4}\sum_{i=1}^{T}\left(\sum_{t=2} x_{it,j}x_{it,l}\right)\left(\sum_{s=1}^{t-1} x_{is,j}x_{is,l}\right) \\ &= V_{j,l}^*.\end{aligned}$$

since all cross terms have zero expectation and $\mathbb{E}\left[(u_{it}^*)^2\right] = 1$. The central limit theorem for independent random variables and Slutsky's theorem imply

$$\left(\frac{1}{N}V^*\right)^{-1/2}\left(\frac{1}{\sqrt{N}}s^*\right) \xrightarrow{d} \mathcal{N}\left(0, I_K\right)$$

and the result follows.

Proof of Theorem 2.4

Using the arguments in Theorem 2.3, LM_{opg} is asymptotically χ_K^2. Regarding the (k,l) element of the covariance matrix of $\tilde{s}_i^*$, note that

$$\begin{aligned}&\left(\frac{1}{\sigma^4}\right)^2\sum_i \mathbb{E}\left[\left(\sum_t x_{it,k}u_{it}\left(\sum_{s=1}^{t-1} x_{is,k}u_{is}\right)\right)\left(\sum_t x_{it,l}u_{it}\left(\sum_{s=1}^{t-1} x_{is,l}u_{is}\right)\right)\middle| X\right] \\ &= \frac{1}{\sigma^4}\sum_i\left(\sum_t x_{it,k}x_{it,l}\right)\left(\sum_{s=1}^{t-1} x_{is,k}x_{is,l}\right).\end{aligned}$$

Next let

$$z_{it,k} = \sigma^{-4}x_{it,k}\sum_{s=1}^{t-1} u_{is}x_{is,k},$$

and notice that

$$\begin{aligned}&\mathbb{E}\left[\sum_i\sum_t u_{it}^2 z_{it,k}z_{it,l}'\middle| X\right] \\ =\ &\left(\frac{1}{\sigma^4}\right)^2\mathbb{E}\left[x_{it,k}u_{it}\left(\sum_{s=1}^{t-1} x_{is,k}u_{is}\right)x_{it,l}u_{it}\left(\sum_{s=1}^{t-1} x_{is,l}u_{is}\right)\middle| X\right] \\ =\ &\frac{1}{\sigma^4}\sum_i\left(\sum_t x_{it,k}x_{it,l}\right)\left(\sum_{s=1}^{t-1} x_{is,k}x_{is,l}\right).\end{aligned}$$

Furthermore

$$\frac{1}{N}\sum_i\sum_t \widetilde{u}_{it}^2\widetilde{z}_{it,k}\widetilde{z}_{it,l} - \frac{1}{N}\sum_i\sum_t u_{it}^2 z_{it,k}z_{it,l} \xrightarrow{p} 0.$$

Since LM_{opg} and LM_{reg} differ only in their variances matrices which vanishes asymptotically,

the result follows.

Proof of Theorem 2.5

As in Honda (1985) the proof of the theorem proceeds in three steps: **(i)** first we show that $\widetilde{\sigma}^2$ remains consistent under the local alternative; **(ii)** second, we incorporate the local alternative into the score vector and **(iii)** establish the asymptotic distribution of the LM statistic.
(i) Note first that with $M_X = I_{NT} - X\left(X'X\right)^{-1}X'$

$$\widetilde{u} = M_X u = M_X\left(D_X v + \epsilon\right),$$

where

$$D_X = \begin{bmatrix} X_1 & & & 0 \\ & X_2 & & \\ & & \ddots & \\ 0 & & & X_N \end{bmatrix}.$$

Hence,

$$\frac{\widetilde{u}'\widetilde{u}}{NT} = \frac{1}{NT}\left(\epsilon'\epsilon - \epsilon' P_X \epsilon + v' D_X' M_X D_X v + v' D_X' M_X \epsilon + v' M_X D_X \epsilon\right).$$

Using Assumptions 2.3 and 2.2, it is straightforward to show that

$$\begin{aligned} \epsilon' P_X \epsilon &= o_p\left(N\right), \\ v' D_X' M_X D_X v &= O_p\left(\sqrt{N}\right), \\ v' D_X' M_X \epsilon &= o_p\left(N\right). \end{aligned}$$

and, thus,

$$\widetilde{\sigma}^2 = \sigma^2 + o_p\left(1\right).$$

(ii) Since $u_i = X_i v_i + \epsilon_i$ and

$$\widetilde{u}_i = X_i v_i + \epsilon_i - X_i\left(\widetilde{\beta} - \beta\right),$$

we obtain

$$\begin{aligned} \frac{1}{\sqrt{N}}\widetilde{s}_k &= \frac{1}{\sqrt{N}}\left(\frac{\sigma^4}{\widetilde{\sigma}^4}\right)\frac{1}{2\sigma^4}\sum_i \epsilon_i' A_i^{(k)} \epsilon_i \\ &\quad + \frac{1}{\sqrt{N}}\left(\frac{\sigma^4}{\widetilde{\sigma}^4}\right)\frac{1}{2\sigma^4}\sum_i v_i'\left(X_i' A_i^{(k)} X_i\right) v_i + o_p(1), \end{aligned} \tag{2.52}$$

for $k = 1, ..., K$, where the order of the remainder term follows by similar arguments as in lemma 2.2.

(iii) Using the same arguments as in the proof of Theorem 2.1, the first term of $\widetilde{s}/\sqrt{N}$ in (2.52) is asymptotically normally distributed. Regarding the second term

$$\frac{1}{\sqrt{N}}\sum_i v_i'\left(X_i' A_i^{(k)} X_i\right) v_i = \frac{1}{N}\sum_i \left(N^{1/4} v_i\right)'\left(X_i' A_i^{(k)} X_i\right)\left(N^{1/4} v_i\right),$$

and by standard results for quadratic forms,

$$\mathbb{E}\left[\left(N^{1/4}v_i\right)'\left(X_iA_i^{(k)}X_i\right)\left(N^{1/4}v_i\right)\middle|X\right]=\text{tr}\left[\left(X_iA_i^{(k)}X_i\right)D_c\right].$$

with

$$D_c=\begin{bmatrix}c_1 & 0 & \cdots & 0\\ 0 & c_2 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & c_K\end{bmatrix}.$$

Thus by the law of large numbers for sums of independent random variables,

$$\frac{1}{2\sqrt{N}\sigma^4}\sum_i v_i'\left(X_i'A_i^{(k)}X_i\right)v_i\xrightarrow{p}\lim_{N\to\infty}\frac{1}{2\sigma^4N}\sum_i\text{tr}\left[\left(X_i'A_i^{(k)}X_i\right)D_c\right].$$

Now

$$\sum_i\text{tr}\Big[\left(X_i'A_i^{(k)}X_i\right)D_c\Big]=\sum_{l=1}^K c_l\left(\sum_i\left(\sum_t x_{it,k}x_{it,l}\right)^2-\frac{1}{NT}\left(\sum_i\sum_t x_{it,k}^2\right)\left(\sum_i\sum_t x_{it,l}^2\right)\right).$$

Define the $K\times 1$ vector ψ elementwise by

$$\psi_k\equiv\sum_{l=1}^K c_l\operatorname*{plim}_{N\to\infty}\left(\frac{1}{N}\sum_i\left(\sum_t x_{it,k}x_{it,l}\right)^2-\frac{1}{T}\left(\frac{1}{N}\sum_i\sum_t x_{it,k}^2\right)\left(\frac{1}{N}\sum_i\sum_t x_{it,l}^2\right)\right).$$

By Slutsky's theorem theorem we obtain

$$\left(\frac{1}{N}\widetilde{V}\right)^{-1/2}\left(\frac{1}{\sqrt{N}}\widetilde{s}\right)\xrightarrow{d}\mathcal{N}\left(\psi,I_K\right),$$

and the theorem follows by the definition of the non-central χ^2 distributed random variable, using $\psi=\Psi c$, where $c=(c_1,\ldots,c_K)'$.

Proof of Theorem 2.6

The proof is analogous to the proof of Theorem 2.5. To show that $\widetilde{\sigma}^2$ remains consistent under the sequence of alternatives we note that

$$\begin{aligned}\epsilon'P_X\epsilon &= O_p\left(1\right),\\ v'D_X'M_XD_Xv &= O_p\left(N^{1/2}T\right),\\ v'D_X'M_X\epsilon/NT &= O_p\left(T^{-1/2}\right)+O_p\left(N^{-1}T^{-1/2}\right).\end{aligned}$$

Using the same arguments as in the proof of Theorem 2.2, $\widetilde{s}/T\sqrt{N}$ has a limiting normal distribution with nonzero mean which is determined by applying the law of large numbers to the second term in (2.52) with proper normalization.

Proof of Theorem 2.7

With the Swamy statistic as described in the text, the proof follows the steps outlined in Appendix A.6 in PY.

Details for Remark 2.7

We study the (k,l) element of $\left(N^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\widetilde{u}_{it}^{2}\widetilde{z}_{it}\widetilde{z}_{it}'\right)$ under the sequence of alternatives in Theorem 2.5. Note that

$$\widetilde{u}_{it}^{2}=\left(\epsilon_{it}+x_{it}'v_{i}\right)^{2}+\left(\widetilde{\beta}-\beta\right)'x_{it}x_{it}'\left(\widetilde{\beta}-\beta\right)-2\left(\epsilon_{it}+x_{it}'v_{i}\right)x_{it}'\left(\widetilde{\beta}-\beta\right), \tag{2.53}$$

and

$$\widetilde{z}_{it,k}=\frac{1}{\widetilde{\sigma}^{4}}x_{it,k}\sum_{s=1}^{t-1}\left(\epsilon_{is}+x_{is}'v_{i}+x_{is}'\left(\widetilde{\beta}-\beta\right)\right)x_{is,k}.$$

implying,

$$\begin{aligned}
&\left(\widetilde{\sigma}^{4}\right)^{2}\widetilde{z}_{it,k}\widetilde{z}_{it,l}=x_{it,k}x_{it,l}\left(\sum_{s=1}^{t-1}\epsilon_{is}x_{is,k}\right)\left(\sum_{s=1}^{t-1}\epsilon_{is}x_{is,l}\right)\\
&+v_{i}'\left(x_{it,k}\sum_{s=1}^{t-1}x_{is}x_{is,k}\right)\left(x_{it,l}\sum_{s=1}^{t-1}x_{is}'x_{is,l}\right)v_{i}\\
&+\left(\widetilde{\beta}-\beta\right)'\left(x_{it,k}\left(\sum_{s=1}^{t-1}x_{is}x_{is,k}\right)\right)\left(x_{it,l}\left(\sum_{s=1}^{t-1}x_{is}'x_{is,l}\right)\right)\left(\widetilde{\beta}-\beta\right)\\
&+x_{it,k}\left(\sum_{s=1}^{t-1}\epsilon_{is}x_{is,k}\right)\left(x_{it,l}\sum_{s=1}^{t-1}\left(x_{is}'v_{i}\right)x_{is,l}\right)+\left(x_{it,k}\sum_{s=1}^{t-1}\left(x_{is}'v_{i}\right)x_{is,k}\right)\left(x_{it,l}\sum_{s=1}^{t-1}\epsilon_{is}x_{is,l}\right)\\
&-x_{it,k}x_{it,l}\left(\sum_{s=1}^{t-1}\epsilon_{is}x_{is,k}\right)\left(\sum_{s=1}^{t-1}x_{is,l}x_{is}'\right)\left(\widetilde{\beta}-\beta\right)\\
&-x_{it,k}x_{it,l}\left(\sum_{s=1}^{t-1}\left(x_{is}'v_{i}\right)x_{is,k}\right)\left(\sum_{s=1}^{t-1}x_{is,l}x_{is}'\right)\left(\widetilde{\beta}-\beta\right)\\
&-x_{it,k}x_{is,l}\left(\sum_{s=1}^{t-1}x_{is}'\left(\widetilde{\beta}-\beta\right)x_{is,k}\right)\left(\sum_{s=1}^{t-1}\epsilon_{is}x_{is,l}\right)\\
&-x_{it,k}x_{is,l}\left(\sum_{s=1}^{t-1}x_{is}'\left(\widetilde{\beta}-\beta\right)x_{is,k}\right)\left(\sum_{s=1}^{t-1}x_{is}'v_{i}x_{is,l}\right).
\end{aligned}\tag{2.54}$$

First, from the first term on the right hand sides of (2.53) and (2.54), we obtain

$$\left(\frac{1}{\widetilde{\sigma}^{4}}\right)^{2}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T}\epsilon_{it}^{2}x_{it,k}x_{it,l}\left(\sum_{s=1}^{t-1}\epsilon_{is}x_{is,k}\right)\left(\sum_{s=1}^{t-1}\epsilon_{is}x_{is,l}\right).$$

Notice that this term has the same probability limit as $\widetilde{V}_{k,l}^{*}/N$, which is equal to $\Psi_{k,l}^{*}$. Next, from the first term on the right-hand side in (2.53) and the second term on the right-hand side

in (2.54),

$$\left(\frac{1}{\widetilde{\sigma}^4}\right)^2 \frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T}\epsilon_{it}^2 v_i'\left(x_{it,k}\sum_{s=1}^{t-1}x_{is}x_{is,k}\right)\left(x_{it,l}\sum_{s=1}^{t-1}x_{it}'x_{is,l}\right)v_i$$
$$=\left(\frac{1}{\widetilde{\sigma}^4}\right)^2 \frac{1}{N^{1.5}}\sum_{i=1}^{N}\sum_{t=2}^{T}\epsilon_{it}^2 \left(N^{1/4}v_i\right)'\left(x_{it,k}\sum_{s=1}^{t-1}x_{is}x_{is,k}\right)\left(x_{it,l}\sum_{s=1}^{t-1}x_{it}'x_{is,l}\right)\left(N^{1/4}v_i\right)$$

Since ϵ_{it} and v_i are independent conditional on X,

$$\mathbb{E}\left[\epsilon_{it}^2 \left(N^{1/4}v_i\right)'\left(x_{it,k}\sum_{s=1}^{t-1}x_{is}x_{is,k}\right)\left(x_{it,l}\sum_{s=1}^{t-1}x_{it}'x_{is,l}\right)\left(N^{1/4}v_i\right)\,\middle|\, X\right]=\sigma^2\text{tr}\left[B_{it}^X D_c\right]$$

with the $K\times K$ matrix $B_{it}^X=\left(x_{it,k}\sum_{s=1}^{t-1}x_{is}x_{is,k}\right)\left(x_{it,l}\sum_{s=1}^{t-1}x_{it}'x_{is,l}\right)$ such that

$$\frac{1}{N^{1.5}}\sum_{i=1}^{N}\sum_{t=2}^{T}\epsilon_{it}^2 \left(N^{1/4}v_i\right)'\left(x_{it,k}\sum_{s=1}^{t-1}x_{is}x_{is,k}\right)\left(x_{it,l}\sum_{s=1}^{t-1}x_{it}'x_{is,l}\right)\left(N^{1/4}v_i\right)=O_p\left(N^{-1/2}\right)$$

Using the properties of ϵ_{it}, v_i and the fact that $\left(\widetilde{\beta}-\beta\right)=o_p\left(1\right)$, it can be shown in a similar manner that all of the remaining terms are of lower order.

2.B Tables

Table 2.1: Monte Carlo experiments with normally distributed errors

		A) Size				B) Power			
		$\widetilde{\Delta}_{\text{adj}}$	CLM	LM	LM_{reg}	$\widetilde{\Delta}_{\text{adj}}$	CLM	LM	LM_{reg}
$T = 10$									
	$N = 10$	6.3	11.8	2.6	4.7	5.5	5.4	11.3	4.1
	$N = 20$	5.6	12.0	3.2	4.3	8.4	4.2	24.8	7.0
	$N = 30$	5.5	11.4	3.7	4.3	11.5	6.3	35.2	10.4
	$N = 50$	5.2	8.9	3.7	4.1	18.9	15.9	51.2	19.6
	$N = 100$	5.4	8.3	4.7	4.8	36.1	46.5	77.5	47.0
	$N = 200$	4.6	7.5	4.9	4.6	65.6	87.3	96.9	83.1
$T = 20$									
	$N = 10$	5.3	15.1	2.6	6.3	17.1	3.4	28.1	12.4
	$N = 20$	5.7	14.2	3.4	5.7	35.0	9.5	53.0	25.8
	$N = 30$	5.9	12.8	3.8	5.8	50.7	23.4	70.5	43.2
	$N = 50$	5.1	10.8	4.1	5.3	74.6	53.5	88.9	71.8
	$N = 100$	4.5	8.3	4.3	4.7	95.7	90.1	99.1	96.5
	$N = 200$	5.1	7.1	5.2	5.5	99.9	98.8	100.0	100.0
$T = 30$									
	$N = 10$	4.7	15.6	2.4	7.0	34.4	4.6	43.0	22.4
	$N = 20$	4.5	14.5	3.5	6.2	64.8	19.7	74.3	50.9
	$N = 30$	5.2	13.2	3.9	5.6	81.9	42.5	88.9	73.3
	$N = 50$	5.3	11.6	4.5	5.9	96.3	76.9	98.2	94.5
	$N = 100$	5.2	8.9	4.3	4.7	100.0	96.0	100.0	99.9
	$N = 200$	5.2	7.4	5.0	5.8	100.0	99.3	100.0	100.0

Note: Rejection frequencies (in %) under the null (panel A) and the alternative hypothesis (panel B). Nominal size is 5%. The model is given in (2.36) for $K = 3$.

Table 2.2: Monte Carlo experiments for variations of the standard design

	A) Size		B) Power			
	LM	LM_{reg}	$\widetilde{\Delta}_{\text{adj}}$	CLM	LM	LM_{reg}
$T = 10$						
$N = 10$	4.5	4.4	8.5	3.8	22.1	6.8
$N = 20$	6.3	4.0	18.5	8.0	43.1	13.4
$N = 30$	7.1	3.8	27.1	19.0	58.1	23.8
$N = 50$	8.5	4.9	45.7	45.7	79.5	46.8
$N = 100$	11.9	8.1	77.8	86.5	96.8	85.0
$N = 200$	17.5	13.6	97.3	98.8	99.9	99.2
$T = 20$						
$N = 10$	3.4	5.5	39.2	6.4	51.9	25.4
$N = 20$	5.0	4.9	70.9	30.8	81.2	57.1
$N = 30$	5.3	4.9	87.6	59.5	93.2	79.5
$N = 50$	6.2	4.7	98.4	88.6	99.6	97.2
$N = 100$	7.3	4.8	100.0	98.8	100.0	100.0
$N = 200$	8.5	5.8	100.0	99.8	100.0	100.0
$T = 30$						
$N = 10$	2.8	6.8	66.8	11.4	71.2	47.1
$N = 20$	4.5	5.5	93.8	52.3	94.4	85.2
$N = 30$	4.9	5.3	98.9	80.1	99.1	96.9
$N = 50$	5.8	5.2	100.0	95.4	100.0	99.9
$N = 100$	5.9	4.4	100.0	99.2	100.0	100.0
$N = 200$	6.8	5.5	100.0	99.9	100.0	100.0

Note: Left panel: rejection frequencies (in %) under the null hypothesis with same design as in Table 2.1 and within transformation to eliminate fixed effects. Right panel: rejection frequencies (in %) under the alternative hypothesis in model (2.36) when $\sigma^2_{ix,k}$ is distributed as χ^2 with two degrees of freedom, see (2.40).

Table 2.3: Monte Carlo experiments with t-distributed errors

A) Size		$\widetilde{\Delta}_{\text{adj}}$	CLM	LM	LM_{adj}	LM^*	LM_{reg}
$T = 10$							
	$N = 10$	6.3	9.0	3.5	2.9	4.0	3.9
	$N = 20$	6.1	10.5	5.4	4.0	6.1	4.5
	$N = 30$	5.6	10.2	6.9	5.0	6.2	4.9
	$N = 50$	5.1	8.9	7.5	5.3	6.6	4.4
	$N = 100$	5.0	7.5	9.8	6.2	7.2	4.9
	$N = 200$	5.5	6.9	10.7	5.7	7.6	5.2
$T = 20$							
	$N = 10$	5.5	13.2	3.5	2.8	3.6	6.0
	$N = 20$	5.8	13.0	5.1	4.0	4.6	6.3
	$N = 30$	5.4	11.7	5.5	4.3	4.8	5.6
	$N = 50$	5.2	10.4	6.7	4.8	5.1	5.3
	$N = 100$	4.9	8.0	8.0	5.5	5.9	5.1
	$N = 200$	5.1	7.0	8.9	5.5	5.4	4.8
$T = 30$							
	$N = 10$	5.8	15.0	3.0	2.6	3.0	7.0
	$N = 20$	5.0	13.6	4.5	3.5	3.9	5.7
	$N = 30$	4.7	12.4	5.2	4.3	4.2	5.5
	$N = 50$	5.1	11.3	5.9	4.7	5.1	5.9
	$N = 100$	5.0	8.5	6.4	5.0	4.8	5.1
	$N = 200$	5.3	7.7	7.3	5.0	5.2	4.9

B) Power		$\widetilde{\Delta}_{\text{adj}}$	CLM	LM	LM_{adj}	LM^*	LM_{reg}
$T = 10$							
	$N = 10$	5.9	3.4	12.4	10.9	11.8	4.9
	$N = 20$	9.7	3.9	25.6	22.7	22.9	7.5
	$N = 30$	13.6	6.9	36.1	31.8	32.4	11.9
	$N = 50$	24.1	16.0	52.4	46.7	47.9	22.6
	$N = 100$	45.3	44.4	78.9	71.8	73.9	49.4
	$N = 200$	76.1	83.1	95.7	92.6	93.8	83.2
$T = 20$							
	$N = 10$	20.0	3.2	30.5	28.6	29.4	13.7
	$N = 20$	39.5	10.1	53.6	50.0	52.5	29.7
	$N = 30$	57.3	22.8	70.5	67.3	68.5	46.7
	$N = 50$	80.0	51.5	88.3	85.5	86.7	72.2
	$N = 100$	97.8	89.5	99.0	98.4	99.0	96.4
	$N = 200$	100.0	98.8	100.0	100.0	100.0	100.0
$T = 30$							
	$N = 10$	37.5	3.7	45.8	43.8	45.1	25.4
	$N = 20$	67.7	19.6	74.3	71.9	73.3	54.3
	$N = 30$	85.4	43.0	88.6	86.6	87.4	74.4
	$N = 50$	97.3	76.2	98.0	97.2	97.7	93.8
	$N = 100$	100.0	95.9	100.0	100.0	100.0	99.9
	$N = 200$	100.0	99.4	100.0	100.0	100.0	100.0

Note: Rejection frequencies (in %) under the null (panel A) and the alternative hypothesis (panel B) with the same design as in Table 2.1 but ϵ_{it} is drawn from a t-distribution with five degrees of freedom, scaled to have unit variance.

Table 2.4: Monte Carlo experiments with χ^2 distributed errors

A) Size		$\widetilde{\Delta}_{\text{adj}}$	CLM	LM	LM_{adj}	LM^*	LM_{reg}
$T = 10$							
	$N = 10$	6.8	6.8	4.6	3.1	5.1	3.4
	$N = 20$	5.4	8.8	7.3	4.6	6.7	4.3
	$N = 30$	5.4	9.6	7.8	4.8	7.0	3.7
	$N = 50$	5.1	8.8	9.4	5.6	7.0	4.5
	$N = 100$	5.3	8.0	12.3	6.5	7.7	4.6
	$N = 200$	4.9	6.9	13.6	6.4	7.7	3.9
$T = 20$							
	$N = 10$	4.9	11.3	4.2	3.1	4.2	5.1
	$N = 20$	5.3	12.6	5.4	3.9	5.3	5.2
	$N = 30$	5.4	12.0	6.8	4.4	4.8	4.9
	$N = 50$	4.9	9.0	7.2	4.6	5.4	5.1
	$N = 100$	5.3	8.6	9.3	5.5	6.3	5.2
	$N = 200$	4.9	7.2	9.6	5.3	6.1	4.9
$T = 30$							
	$N = 10$	5.2	13.6	3.7	2.9	3.4	6.5
	$N = 20$	5.2	13.2	4.7	3.5	3.9	5.7
	$N = 30$	4.9	12.7	6.1	4.4	4.8	5.5
	$N = 50$	4.9	10.3	6.1	4.4	5.4	4.8
	$N = 100$	5.0	9.5	7.6	5.2	5.4	5.4
	$N = 200$	4.8	7.1	9.2	5.6	6.2	5.2

B) Power		$\widetilde{\Delta}_{\text{adj}}$	CLM	LM	LM_{adj}	LM^*	LM_{reg}
$T = 10$							
	$N = 10$	6.6	3.2	13.8	11.5	13.2	4.9
	$N = 20$	11.3	4.3	27.8	22.2	23.9	8.8
	$N = 30$	17.9	6.7	36.4	29.4	32.1	12.8
	$N = 50$	29.2	14.7	53.6	43.2	47.7	23.6
	$N = 100$	58.7	43.1	78.4	68.7	72.2	50.4
	$N = 200$	87.8	81.4	95.5	91.7	93.7	83.1
$T = 20$							
	$N = 10$	21.5	3.3	29.8	26.1	28.4	15.4
	$N = 20$	44.2	10.7	56.0	51.0	54.5	31.6
	$N = 30$	62.7	23.5	71.5	66.3	69.3	49.4
	$N = 50$	85.4	51.9	88.9	85.3	87.7	74.9
	$N = 100$	98.9	89.5	99.3	98.8	98.9	96.2
	$N = 200$	100.0	98.7	100.0	100.0	100.0	99.9
$T = 30$							
	$N = 10$	39.4	4.5	44.1	41.3	43.6	27.2
	$N = 20$	71.1	20.7	73.7	70.6	73.0	55.3
	$N = 30$	88.4	42.0	89.0	86.6	88.2	75.9
	$N = 50$	98.1	75.1	97.9	97.1	97.8	94.1
	$N = 100$	100.0	96.0	100.0	100.0	100.0	99.9
	$N = 200$	100.0	99.2	100.0	100.0	100.0	100.0

Note: Rejection frequencies (in %) under the null (panel A) and the alternative hypothesis (panel B) with the same design as in Table 2.1 but ϵ_{it} is drawn from a χ^2-distribution with two degrees of freedom, centered at zero and scaled to have unit variance.

Table 2.5: Monte Carlo experiments with non-diagonal matrix Σ_v

Power		$\widetilde{\Delta}_{\text{adj}}$	CLM	LM	LM_{reg}
$T = 10$					
	$N = 10$	5.6	5.3	11.4	4.4
	$N = 20$	8.8	3.8	24.7	6.6
	$N = 30$	11.5	6.5	34.7	10.4
	$N = 50$	18.5	15.7	50.5	19.5
	$N = 100$	35.7	46.3	77.8	46.4
	$N = 200$	65.3	86.6	96.8	83.2
$T = 20$					
	$N = 10$	17.5	3.4	28.3	11.7
	$N = 20$	34.3	9.3	53.6	25.8
	$N = 30$	50.0	22.8	70.0	42.4
	$N = 50$	74.1	51.8	88.8	71.6
	$N = 100$	95.3	89.7	99.2	96.6
	$N = 200$	99.9	98.8	100.0	100.0
$T = 30$					
	$N = 10$	34.1	4.6	42.6	21.8
	$N = 20$	64.3	19.5	74.1	50.4
	$N = 30$	81.5	42.5	88.4	73.0
	$N = 50$	96.1	76.1	98.3	94.2
	$N = 100$	100.0	96.0	100.0	99.9
	$N = 200$	100.0	99.3	100.0	100.0

Note: Rejection frequencies (in %) under the alternative hypothesis with the same design as in Table 2.1 but Σ_v being no longer diagonal, see section 2.6.4.

Chapter 3

Predictive regressions with possibly persistent regressors under asymmetric loss

3.1 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 squared error loss. 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.[2] In addition, Capistrán (2008) even finds evidence of time-varying asymmetric preferences. More recently,

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

[2] See also Clements et al. (2007).

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 directly to estimation under the relevant loss function (see Granger (1969) and Weiss (1996)); that is, one should estimate relevant parameters by minimizing the observed loss to obtain an estimate of the forecast optimal under that loss function. 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)). Suppose a regression disturbance term has zero conditional mean given the regressor, yet is conditionally heteroskedastic. Under the mean squared error criterion, 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.[3] 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.

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 endogeneous 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. For instance, 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.[4]

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

[3] See Christoffersen and Diebold (1997), for example.

[4] See also Stambaugh (1999) and Phillips and Lee (2013) for a recent review of inference in predictive regressions.

integrated regressors, for which the mean reversion parameter cannot be consistently estimated. Exclusively for nearly integrated regressors, Jansson and Moreira (2006) propose a conditional likelihood approach based on a sufficient statistic for the local-to-unity parameter. Alternatively, Hjalmarsson (2010) uses panel data to reduce the effects of endogeneity, again, in an OLS setup. 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, investigates the predictive ability of economic variables that are persistent. 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 given in Section 3.2.1), 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) argue however that 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.[5] To allow for these different possibilites, we consider a potential predictor to be persistent if the regressor, suitably normalized, converges weakly to a continuous-time process.

To provide correct inference, we draw on an instrumental variable (IV) approach as studied by Breitung and Demetrescu (2013) and propose in Section 3.2.2 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 integration of the predictor.

We then reexamine the well-known forward premium puzzle in Section 3.4. Evidence for de-

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

viations from MSE loss is presented for a collection of exchange rates, and 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.

Before proceeding to the main part of this chapter, 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 or backshift 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 cadlag 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.

3.2 Estimation and inference under asymmetric loss

Consider the prototypical predictive regression model

$$y_t = \beta_0 + \beta_1 x_{t-1} + u_t, \tag{3.1}$$

where the null of interest is $\beta_1 = 0$. 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,

$$x_t = v_t + \sum_{j=1}^{t} \psi_{j,T} v_{t-j}. \tag{3.2}$$

The shocks u_t and v_t are taken to satisfy standard regularity conditions, see Assumption 3.2 below; in particular they are allowed to be contemporaneously dependent at time t to capture endogeneity in the predictive regression model (3.1). 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.[6] See Example 3.1 below. At the same time, short-run dynamics is allowed for; e.g. in the near-integrated case by letting $\psi_{j,T}$

[6]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.

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 3.1 *A process x_t is highly persistent in our framework if the coefficients $\psi_{j,T}$ in (3.2) 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*

$$\frac{1}{n_T} x_{[sT]} \Rightarrow \sigma_v X(s), \tag{3.3}$$

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 3.2 *A process x_t is weakly persistent in our framework if the coefficients $\psi_{j,T}$ in (3.2) are of such nature that*

$$\lim_{T\to\infty} \sum_{j=1}^{T} \psi_{j,T}^2 = C < \infty. \tag{3.4}$$

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 (3.3) or (3.4) holds.

Example 3.1 *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 3.1 *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).[7] The sign-based test proposed by Campbell and Dufour (Campbell and Dufour) 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 3.1 makes $\mathcal{L}$ strictly convex and smooth with first-order derivative given by

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

[7] 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 pth derivative is discontinuous, and the $p-1$st derivative satisfies a uniform Lipschitz condition.

The natural choice when considering inference under an asymmetric loss function is to base predictability tests on estimation of (3.1),

$$y_t = \widehat{\beta}_0 + \widehat{\beta}_1 x_{t-1} + \widehat{u}_t, \tag{3.5}$$

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

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right)' = \underset{\left(\beta_0^*,\beta_1^*\right)'}{\arg\min} \sum_{t=2}^{T} \mathcal{L}\left(y_t - \beta_0^* - \beta_1^* x_{t-1}\right). \tag{3.6}$$

Note that the intercept $\widehat{\beta}_0$ in (3.5) 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 3.2 below), 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

$$\beta_0 = \underset{\beta_0^*}{\arg\min}\ \mathrm{E}\left(\mathcal{L}\left(u_t - \beta_0^*\right)\right). \tag{3.7}$$

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.

Assumption 3.2 *The series y_t and x_t, $t = 1, \ldots, T$, are generated as in* (3.1) *and* (3.2) *such that either* (3.3) *(high persistence) or* (3.4) *(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.*

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

$$t_{\beta_1} = \frac{\widehat{\beta}_1}{s.e.\left(\widehat{\beta}_1\right)}, \tag{3.8}$$

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

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

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).
We show in the following, however, that the limiting behavior of $\widehat{\beta}$ and of t_{β_1} is nonstandard if x_t is highly persistent whenever there is endogeneity. The actual distribution depends on the limit of x_t (and is e.g. different when x_t is fractionally or near integrated). Moreover, what endogeneity stands for is loss-function specific; in other words, if u_t and v_t are uncorrelated, the usual OLS estimator is not biased and its *t* statistic has a standard normal distribution, but, unless u_t and v_t are independent, one cannot guarantee that estimation under the relevant loss function leads to a standard normal *t* statistic.

3.2.1 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

$$\widetilde{u}_t = \mathcal{L}^{(1)}\left(u_t - \beta_0\right), \tag{3.9}$$

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 3.2,

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

jointly with (3.3) 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 3.1 *Under Assumptions 3.1 and 3.2, as $T \to \infty$,*

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

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 3.2 *Under the assumptions of theorem 3.1, as $T \to \infty$,*

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

Remark 3.1 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 3.2 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 3.3 The quantile regression result of Lee (2012) for near integration can formally be derived from theorem 3.2 using the method of Phillips (1991) and letting X be an OU process.

Remark 3.4 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. Existing solutions are typically suggested for near-integrated regressors, or for a particular $\mathcal{L}$ (OLS or LAD). The following subsection considers a simple solution which still has power.

3.2.2 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 3.2. A simple way out is variable addition 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.[8] To avoid such a loss, we follow Breitung and Demetrescu (2013) and resort to overidentified estimation and testing. We adapt in the following their Anderson-Rubin (AR) type statistic to M estimation and testing under the relevant loss in (3.6).

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

[8] This is a tedious, yet straightforward extension of the results of Section 3.2.1.

the original variable to qualify as a valid instrument. The second is strictly exogenous but persistent. In the nearly integrated framework, several ways to construct such an instrument are offered, including the mildly integrated process $(1-\gamma_T L)_+^{-1}\Delta x_{t-1}$ for $\gamma_T = 1 - a/T^\delta$, with $a > 0$, and $0 < \delta < 1$. 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 to guarantee exogeneity, and the class includes, for example, randomly generated random walks or functions of (scaled) time.

Model (3.2) 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 3.3 (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 predictor. While it is true that Breitung and Demetrescu (2013) allow for a wider class of type-I instruments when x_t is near-integrated, this is the price to pay for having any kind of persistent behavior of x_t.

Assumption 3.3 *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) *as* $T \to \infty$, plim $T^{-1}\sum_{t=2}^{T}\left(1, \mathbf{z}_{t-1,T}'\right)'\left(1, \mathbf{z}_{t-1,T}'\right) = \Sigma_z$, *for a 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).

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, Komunjer, and Timmermann (2005). Concretely, the

predictability test is conducted with

$$\mathcal{T} = \lambda' \Lambda \lambda \tag{3.10}$$

with

$$\lambda = \sum_{t=2}^{T} \widetilde{\mathbf{z}}_{t-1,T} \mathcal{L}^{(1)} \left(y_t - \widehat{\beta}_0 \right)$$

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

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

Its properties under the null are summarized in the following result.

Theorem 3.3 *Under Assumptions (3.1) - (3.3) and the null hypothesis, as $T \to \infty$,*

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

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

$$H_{1,T} : \ \beta_1 = \frac{b}{n_T \sqrt{T}}, \tag{3.11}$$

for highly persistent regressors and

$$H_{1,T} : \ \beta_1 = \frac{b}{\sqrt{T}}, \tag{3.12}$$

for weakly persistent regressors, and obtain the following result.

Theorem 3.4 (i) *If the predictor x_{t-1} is persistent, then under Assumptions 3.1 - 3.3 and the sequence of local alternatives (3.11), 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 3.1 - 3.3 and the sequence of local alternatives (3.12), 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. 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 $\widehat{\beta}_1$), should one be able to fix its size problem in the general setup of Definition 3.1. 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 stationary.

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 regres-

sion system,

$$y_t = \beta x_{t-1} + u_t, \tag{3.13}$$
$$x_t = \rho_T x_{t-1} + v_t,$$

for $t = 1, \ldots, T$, with 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,

$$\mathcal{L}(u_t) = ((1 - 2\alpha)\,\mathbf{1}(u_t < 0) + \alpha)\,|u_t|^2. \tag{3.14}$$

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 endogeneous regression system.

As has been pointed out in section 3.2.1, endogeneity under asymmetric loss is determined by the correlation parameter $\widetilde{\omega} = \mathbb{E}[\widetilde{u}_t v_t]$ which need not coincide with correlation between the disturbance terms in the linear model (3.13). To discuss this distinction, suppose u_t is characterized by multiplicative heteroskedasticity,

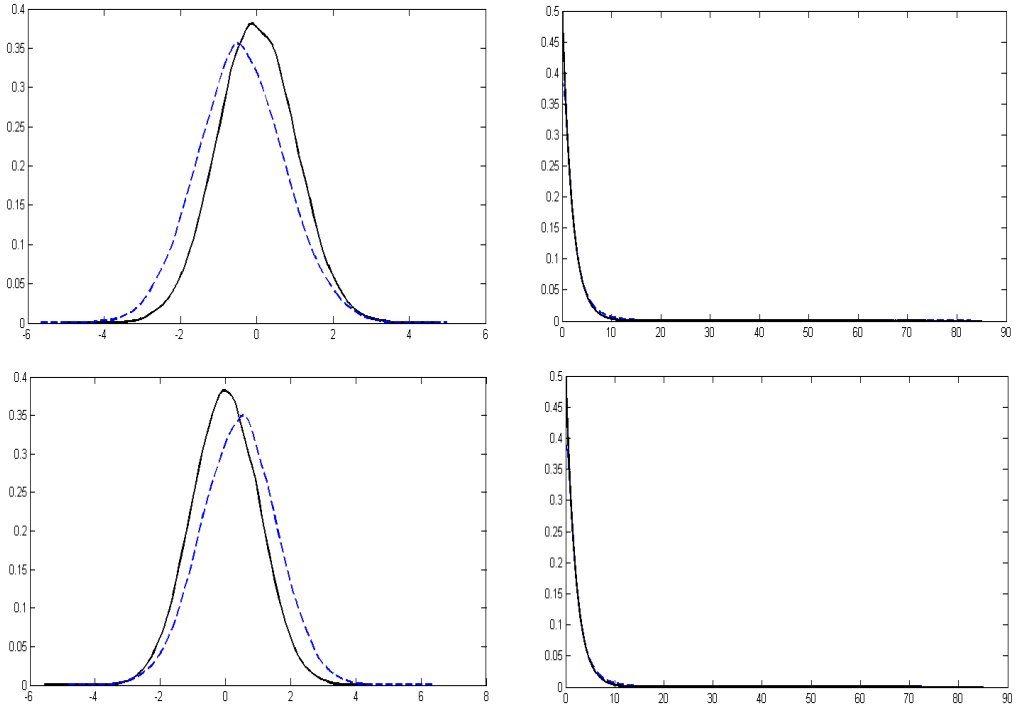
$$u_t = \sigma_t \epsilon_t = \sigma \sqrt{f(v_t)}\epsilon_t. \tag{3.15}$$

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 predictabiliy 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, (3.15) implies $\omega = \mathbb{E}[u_t v_t] = 0$, so even if the predictor of interest is persistent, standard inference applies in (3.13). In contrast, $\widetilde{\omega}$ may differ from zero, which, although analytically intractable, is exemplified for the following choice of $f(\cdot)$

$$f(v_t) = (|v_t| - \gamma v_t)^2, \tag{3.16}$$

Figure 3.1: Densities of t-statistics for loss function parameter $\alpha = 0.2$



Note: The d.g.p. is given in (3.13) with $\rho_T = 1$, $\beta = 0$ and $T = 100$ and $f(v_t)$ as in (3.16). 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).

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 (3.8). 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 3.1 illustrates the consequences of an asymmetric loss function for inference in predictive regressions when $f(v_t)$ is given in (3.16) 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 3.2. 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 a common null hypothesis in a return predictability example 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 provides 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 3.3.

3.4 Robust inference in forward premium regressions

3.4.1 The forward premium puzzle under asymmetric loss

The foreign exchange rate market provides an opportunity to test the rational expectations hypothesis: 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.[9] 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

$$s_{t+k} - s_t = \gamma_0 + \gamma_1 \left(f_t^{t+k} - s_t\right) + u_{t+k}, \tag{3.17}$$

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 idionsyncratic error, see, among others, Fama (1984). If agents are risk neutral, the rational hypothesis corresponds

[9] See Geweke and Feige (1979), for example.

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

$$s_{t+k} - f_t^{t+k} = \beta_0 + \beta_1 \left(f_t^{t+k} - s_t \right) + u_{t+k}, \tag{3.18}$$

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 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 (3.17) 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 (3.17) or (3.18) are estimated taking into account the respective loss function, and hypothesis tests in these models are carried out using these estimates.

In fact, Pierdzioch, Rülke, and Stadtmann (2012a) examine individual exchange rate forecasters and find that the symmetric, quadratic loss function does not apply to all market participants. More recently, Christodoulakis and Mamatzakis (2013) find evidence for the so called Quad-Quad loss function in monthly exchange rate series of G7 countries. This specification corresponds to $p = 2$ and α potentially different from 0.5 in our setup, as in

$$\mathcal{L}\left(s_{t+k} - f_t^{t+k}\right) = \left((1-2\alpha)\,\mathbf{1}\left(s_{t+k} - f_t^{t+k} < 0\right) + \alpha\right) |s_{t+k} - f_t^{t+k}|^2, \tag{3.19}$$

which imposes a higher penalty for overprediction of the exchange rate if $\alpha < 0.5$, while underprediction is more costly if $\alpha > 0.5$. Furthermore, it has been pointed out by Liu and Maynard (2005) or Gospodinov (2009) that the forward premium is a possibly persistent process and shocks to the spot and forward rates are correlated. The standard t statistic to test $\beta_1 = 0$ in (3.18) may thus be biased resulting in unreliable inference. Combining this evidence for

asymmetric loss and persistent regressors allows us to test the rational expectations hypothesis with the inferential methods developed in the previous sections. 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.

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 3.1 presents summary statistics of the spot and forward rates and the relevant regression variables used in (3.18). 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 a persistent predictor in these cases.

Next, we estimate endogeneity in the regression system by

$$\widetilde{\omega} = \frac{\sigma_{\widetilde{u}v}}{\sigma_{\widetilde{u}}\sigma_v}, \tag{3.20}$$

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 autoreggressive model for the forward premium with lag selection by the BIC criterion, while $\widehat{u}_{t+k}$ are the residuals in (3.18). Table 3.2 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. Hence, for comparison, inference is carried out with both the M test based on (3.8) and the AR statistic.

Table 3.1: Summary statistics for 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.

Table 3.2: 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 (3.20). The sample for EUR begins on 01/03/1999.

3.4.2 Estimating loss function parameters

We follow Elliott, Komunjer, and Timmermann (2005) to estimate the paramater α in (3.19). 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,

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

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 satsifies 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 $\widehat{h}=1/\left(T-5\right)\sum_{t=2}^{T-4}w_{t+3}|s_{t+4}-f_t^{t+4}|$. In view of the results of Section 3.2.1, it is however 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 3.1. 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 3.3a and 3.3b present point estimates of the parameter α in (3.19) for the different series using the procedure suggested by Elliott, Komunjer, and Timmermann (2005). The standard errors and probability values for testing the hypothesis $\alpha=0$ against $\alpha\neq0$ 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, Rülke, and Stadtmann (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,

Table 3.3a: 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 3.3b: 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: 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 $\alpha \neq 0.5$.

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 $\alpha = 0.5$.

3.4.3 Inference with the t statistic and the robust 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 (3.18) 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 t 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.

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 3.4 and 3.5 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, however, some further comments can be made in the full sample. First, table 3.2 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 considered to yield valid inference. The null hypothesis is rejected for the symmetric case. Given the evidence for asymmetric loss with an estimated loss function parameter of about 0.55 or larger, the null hypothesis is bareley rejected or not rejected in these cases. Next, considering the results for the AR statistic, for the

Table 3.4: p values for the test of $\beta_1 = 0$ using the 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 (3.8).

Table 3.5: 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 (3.10).

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 3.3a 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 3.3a.

3.5 Concluding remarks

This chapter extends 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 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 stationary or highly persistent. 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 expecations hypothesis.

Appendix to Chapter 3

3.A Proofs

Preliminary results

Lemma 3.1 *Let Assumptions 3.1 and 3.2 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 $1 \leq k \leq p$*

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

5. *Similarly,*

$$\frac{1}{n_T^k T} \sum_{t=21}^{T} x_{t-1}^k \widetilde{u}_t^2 \Rightarrow \sigma_{\widetilde{u}}^2 \sigma_v^k \int_0^1 X^k(s) \, \mathrm{d}s.$$

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*

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

Proof.

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 $j_1, \dots j_{2p}$ 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 3.1.

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

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

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 $x_{t-1}^k \left(\mathcal{L}^{(2)}\left(u_t - \beta_0\right) - \widetilde{\kappa}\right)$ is a martingale difference sequence given the iid property of $\left(u_t, v_t\right)'$ and thus of $\left(\mathcal{L}^{(2)}\left(u_t - \beta_0\right) - \widetilde{\kappa}, v_t\right)'$, so

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

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

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

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

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

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 $\widetilde{\beta}_0^* - \beta_0 = o_p(1)$ 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,

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

Now,

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

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

■

Lemma 3.2 *For $\widehat{\beta}_0$ and $\widehat{\beta}_1$ from* (3.6) *it holds under persistence as $T \to \infty$ that*

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

such that

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

Proof.
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 determninistic function, minimized at the true values of the parameters, then argmin estimators are consistent" is used; see Chapter 4 of Amemyia (1985).
In order to establish the consistency of $\widehat{\beta}_1$, we distinguish two cases.

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

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

   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

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

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

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

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

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

Since $\mathrm{E}(\mathcal{L}(u_t - \beta_0^*))$ 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 impliy uniform convergence Andersen and Gill (1982, Lemma II.1) to the argmin of $\mathrm{E}(\mathcal{L}(u_t - \beta_0^*))$. But the argmin is indeed β_0 according to its definition (3.7), 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

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

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 3.1 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 3.3 *Under the definition of persistence and assumptions 3.1 - 3.3,*

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

as $T \to \infty$.

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

$$\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1} = \sum_{t=2}^{T}\left(\widetilde{S}_{t-1} - \widetilde{S}_{t-2}\right)x_{t-1} = \widetilde{S}_{T-1}x_{T-1} - \sum_{t=2}^{T-1}\widetilde{S}_{t-1}\Delta x_t. \tag{3.21}$$

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

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

under persistence. Regarding the second term on the r.h.s. in (3.21),

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

Now $\sum_{s=2}^{T}\Delta x_s = O_p\left(n_T\right) = \sum_{t=2}^{T-1}\left(t-1\right)/T\Delta x_t$ 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 3.3. 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 3.1.
■

Proofs of the main results

Proof of Theorem 3.1

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

$$\begin{pmatrix} \sum\widetilde{u}_t \\ \sum x_{t-1}\widetilde{u}_t\end{pmatrix} =$$
$$-\begin{pmatrix} \sum\mathcal{L}^{(2)}\left(y_t-\widetilde{\beta}_0-\widetilde{\beta}_1x_{t-1}\right) & \sum x_{t-1}\mathcal{L}^{(2)}\left(y_t-\widetilde{\beta}_0-\widetilde{\beta}_1x_{t-1}\right) \\ \sum x_{t-1}\mathcal{L}^{(2)}\left(y_t-\widetilde{\beta}_0-\widetilde{\beta}_1x_{t-1}\right) & \sum x_{t-1}^2\mathcal{L}^{(2)}\left(y_t-\widetilde{\beta}_0-\widetilde{\beta}_1x_{t-1}\right)\end{pmatrix}\begin{pmatrix} \widehat{\beta}_0-\beta_0 \\ \widehat{\beta}_1-\beta_1\end{pmatrix},$$

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 $\widetilde{\beta}_0 - \beta_0 = o_p\left(1\right)$.
Using Lemma 3.1 item 5, it follows that

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

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

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

or

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

with

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

and

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

with

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

leading with Lemma 3.1 to the desired result.

Proof of Theorem 3.2

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

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

such that, using Lemma 3.1 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 3.1.

Proof of Theorem 3.3

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 $\left(T-1\right) \times 1$ 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,

$$\mathcal{L}^{(1)}\left(u_t - \widehat{\beta}_0\right) = \mathcal{L}^{(1)}\left(u_t - \beta_0\right) - \left(\widehat{\beta}_0 - \beta_0\right)\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t}\right), \tag{3.22}$$

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

$$q_T = \begin{bmatrix} \frac{1}{\sqrt{T}} \sum_{t=2}^{T} z_{t-1,T}^{(I)} \widetilde{u}_t - \sqrt{T} \left(\widehat{\beta}_0 - \beta_0 \right) \frac{1}{T} \sum_{t=2}^{T} z_{t-1,T}^{(I)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) \\ \frac{1}{\sqrt{T}} \sum_{t=2}^{T} z_{t-1,T}^{(II)} \widetilde{u}_t - \sqrt{T} \left(\widehat{\beta}_0 - \beta_0 \right) \frac{1}{T} \sum_{t=2}^{T} z_{t-1,T}^{(II)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) \end{bmatrix}. \tag{3.23}$$

Note that from (3.22) 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

$$\sqrt{T} \left(\widehat{\beta}_0 - \beta_0 \right) = \frac{\frac{1}{\sqrt{T}} \sum_{t=2}^{T} \widetilde{u}_t}{\frac{1}{T} \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right)}. \tag{3.24}$$

Let

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

with

$$a_{T,13} = \frac{\frac{1}{T} \sum_{t=2}^{T} z_{t-1,T}^{(I)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right)}{\frac{1}{T} \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right)}, \tag{3.25}$$

$$a_{T,23} = \frac{\frac{1}{T} \sum_{t=2}^{T} z_{t-1,T}^{(II)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right)}{\frac{1}{T} \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right)}, \tag{3.26}$$

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 3.2 and 3.3, 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

$$A_T \xrightarrow{p} A, \tag{3.27}$$

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 (3.27), we verify that

$$\frac{1}{T} \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) \xrightarrow{p} \widetilde{\kappa}^{(2)}, \tag{3.28}$$

$$\frac{1}{T} \sum_{t=2}^{T} z_{t-1}^{(I)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) \xrightarrow{p} 0, \tag{3.29}$$

$$\frac{1}{T} \sum_{t=2}^{T} z_{t-1}^{(II)} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) \xrightarrow{p} \widetilde{\kappa}^{(2)} \Sigma_z^{13}. \tag{3.30}$$

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 3.3, implying

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

To establish (3.28)-(3.30), we use arguments similar to those in the proof of Lemma 3.1.6. 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 $u_t - \beta_0$,

$$\begin{aligned} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) = \mathcal{L}^{(2)} (u_t - \beta_0) &+ \sum_{j=3}^{p-2} \frac{1}{(j-2)!} \mathcal{L}^{(j)} (u_t - \beta_0) \left(\beta_0 - \widehat{\beta}_{0,t} \right)^{j-2} \\ &+ \frac{1}{(p-3)!} \mathcal{L}^{(p-1)} \left(u_t - \widehat{\beta}_{0,t}^{*} \right) \left(\beta_0 - \widehat{\beta}_{0,t} \right)^{p-3}. \end{aligned} \tag{3.31}$$

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

$$\begin{aligned} \sum_{t=2}^{T} \mathcal{L}^{(2)} \left(u_t - \widehat{\beta}_{0,t} \right) = \sum_{t=2}^{T} \mathcal{L}^{(2)} (u_t - \beta_0) &+ \sum_{t=2}^{T} \left(\sum_{j=3}^{p-2} \frac{1}{(j-2)!} \mathcal{L}^{(j)} (u_t - \beta_0) \left(\beta_0 - \widehat{\beta}_{0,t} \right)^{j-2} \right) \\ &+ \frac{1}{(p-3)!} \sum_{t=2}^{T} \mathcal{L}^{(p-1)} \left(u_t - \widehat{\beta}_{0,t}^{*} \right) \left(\beta_0 - \widehat{\beta}_{0,t} \right)^{p-3}. \end{aligned} \tag{3.32}$$

As a consequence of assumption 3.2, $\frac{1}{T} \sum_{t=2}^{T} \mathcal{L}^{(2)} (u_t - \beta_0) \xrightarrow{p} \widetilde{\kappa}^{(2)}$, 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 (3.32). 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 (3.28) holds.
Turning to (3.29),

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

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 3.2 and 3.3 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 3.2, 3.3, 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 3.3,

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

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 3.1.
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 (3.29).
Exactly analogous arguments apply to (3.30). 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 (3.30), completing the first part of the proof.
Regarding $(\mathbf{ii})$, let $\mathcal{Q}_{ij}$ denote the (i,j) element of $\mathcal{Q}$. Here,

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

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 3.2 and 3.3, 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 3.2 and 3.3.
Using (3.22),

$$Q_T = \widetilde{Q}_T + R_{1T}^Q + R_{2T}^Q, \tag{3.33}$$

where

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

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

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

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 3.2 and 3.3, $\mathbb{E}\left|\left(z_{t-1,T}^{(I)}\right)^2 \widetilde{u}_t^2\right| < \infty$, 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 3.2 and 3.3. 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 (3.31), 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 (3.24) 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 3.4

The proof follows the steps of the proof of theorem 3.3. 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

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

Therefore, with the scaling matrix D_T as defined in the proof of theorem 3.3,

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

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 3.3. To this end, using a Taylor expansion around $u_t - \beta_0$,

$$\mathcal{L}^{(2)}\left(u_t - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right) = \mathcal{L}^{(2)}\left(u_t - \beta_0\right) \tag{3.34}$$

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

$$+ \frac{1}{(p-3)!}\mathcal{L}^{(p-1)}\left(u_t - \widehat{\beta}_{0,t}^{*} + \frac{\widehat{b}^{*}}{n_T\sqrt{T}} x_{t-1}\right)\left(\beta_0 - \widehat{\beta}_{0,t} + \frac{\widehat{b}}{n_T\sqrt{T}} x_{t-1}\right)^{p-1}. \tag{3.35}$$

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 3.2, 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,

$$\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)\xrightarrow{p}0, \tag{3.36}$$

$$\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(II)}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{n_T\sqrt{T}}x_{t-1}\right)\xrightarrow{p}0. \tag{3.37}$$

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

$$Q_{11}=\sigma_{\widetilde{u}}^2\lim_{T\to\infty}\frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\left[\left(\widetilde{z}_{t-1}^{(I)}\right)^2\right], \tag{3.38}$$

$$Q_{22}=\sigma_{\widetilde{u}}^2\lim_{T\to\infty}\frac{1}{T}\sum_{t=2}^{T}\left(\widetilde{z}_{t-1,T}^{(II)}\right)^2, \tag{3.39}$$

$$Q_{12}=0,$$

where the first two limits are finite by assumption 3.3.
Second, we argue that $\Delta_{1,T}\xrightarrow{p}0$ and

$$\Delta_{2,T}\Rightarrow\Delta_2\equiv\widetilde{\kappa}^{(2)}\,\sigma_v\,\begin{bmatrix}0\\ \int_0^1\widetilde{Z}(s)X\left(s\right)ds\end{bmatrix}, \tag{3.40}$$

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 (3.39), $\Delta_{1,T}\xrightarrow{p}0$, and (3.40) 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 3.1, item 3, and combining this result with (3.36) and (3.37).
Finally, for $\Delta_{2,T}$, consider the first component

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

where we made use of (3.34). 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 3.1. Finally, the third term converges to zero in probability by lemma 3.3.
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 3.3.
(**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

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

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

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

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

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

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,

$$\begin{aligned}\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1}\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_{0,t}+\frac{\widehat{b}}{\sqrt{T}}x_{t-1}\right)&=\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1}\left(\mathcal{L}^{(2)}\left(u_t-\widehat{\beta}_0\right)-\widetilde{\kappa}^{(2)}\right)\\&+\widetilde{\kappa}^{(2)}\frac{1}{T}\sum_{t=2}^{T}\widetilde{z}_{t-1}^{(I)}x_{t-1}+o_p\left(1\right),\end{aligned}\tag{3.41}$$

and the first term can be further decomposed into

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

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 (3.41),

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

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 (3.38) and the properties of the non-central χ^2 distribution.

Bibliography

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

Alper, C., S. Fendoglu, and B. Saltoglu (2012). MIDAS volatility forecast performance under market stress: evidence from emerging stock markets. *Economics Letters 117*, 528–532.

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*, 1100–1120.

Andersen, T. and T. Bollerslev (1998). Answering the sceptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review 39*, 885–905.

Andersen, T., T. Bollerslev, F. Diebold, and P. Labys (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association 96*, 42–55.

Andreou, E., E. Ghysels, and A. Kourtellos (2013). Should macroeconomic forecasters use daily financial data and how? *Journal Business & Economic Statistics 31*, 240–251.

Anup, A. and G. Maddala (1984). Ridge estimators for distributed lag models. *Communications in Statistics - Theory and Methods 13*, 217–225.

Arellano, M. and O. Bover (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics 68*, 29–51.

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

Asgharian, H., A. Hou, and F. Javed (2013). The importance of the macroeconomic variables in forecasting stock return variance: a GARCH-MIDAS approach. *Journal of Forecasting 32*, 600–612.

Baltagi, B., Q. Feng, and C. Kao (2011). Testing for sphericity in a fixed effects panel data model. *The Econometrics Journal 14*, 25–47.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, 307–327.

Breitung, J. and M. Demetrescu (2013). Instrumental variable and variable addition based inference in predictive regressions. Working paper, Department of Economics, University of Bonn.

Breitung, J., S. Elengikal, and C. Roling (2013). Forecasting inflation using daily data: a nonparametric MIDAS approach. Working paper, Department of Economics, University of Bonn.

Breusch, T. S. and A. R. Pagan (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies 47*, 239–253.

Brooks, C. and G. Persand (2003). Volatility forecasting for risk management. *Journal of Forecasting 22*, 1–22.

Campbell, B. and J. M. Dufour. Exact nonparametric orthogonality and random walk tests. *The Review of Economics and Statistics 77*.

Campbell, J. and L. Viceira (2001). Who should buy long-term bonds? *American Economic Review 91*, 99–127.

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

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

Chipman, T. (2011). *Advanced Econometric Theory*. Taylor and Francis.

Christodoulakis, G. and E. Mamatzakis (2008). An assessment of the EU growth forecasts under asymmetric preferences. *Journal of Forecasting 27*, 483–492.

Christodoulakis, G. and E. Mamatzakis (2009). Assessing the prudence of economic forecasts in the EU. *Journal of Applied Econometrics 24*, 583–606.

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

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

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

Clements, M. and A. Galvão (2009). Forecasting US output growth using leading indicators: an appraisal using MIDAS models. *Journal of Applied Econometrics 24*, 1187–1206.

Clements, M., F. Joutz, and H. Stekler (2007). An evaluation of the forecasts of the Federal Reserve: a pooled approach. *Journal of Applied Econometrics 22*, 121–136.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press.

Davidson, J. and P. Sibbertsen (2005). Generating schemes for long memory processes: regimes, aggregation and linearity. *Journal of Econometrics 128*, 253–282.

Dolado, J. and H. Lütkepohl (1996). Making Wald tests work for cointegrated VAR systems. *Econometric Reviews 15*, 369–386.

Elliott, G., I. Komunjer, and A. Timmermann (2005). Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies 72*, 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*, 672–700.

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

Engle, R. (2004). Risk and volatility: econometric models and financial practice. *American Economic Review 94*, 405–420.

Engle, R., E. Gyhsels, and B. Sohn (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics 95*, 776–797.

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

Farebrother, R. (1978). Partitioned ridge regression. *Technometrics 20*, 121–122.

Forsberg, L. and E. Ghysels (2007). Why do absolute returns predict volatility so well? *Journal of Financial Econometrics 5*, 31–67.

Fu, W. (1998). Penalized Regressions: the Bridge versus the Lasso. *Journal of Computational and Graphical Statistics 7*, 397–416.

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*, 334–341.

Ghysels, E., P. Santa Clara, and R. Valkanov (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics 131*, 59–95.

Ghysels, E., A. Sinko, and R. Valkanov (2007). MIDAS regressions: further results and new directions. *Econometric Reviews 26*, 53–90.

Ghysels, E. and R. Valkanov (2012). Forecasting volatility with MIDAS. In L. Bauwens, C. Hafner, and S. Laurent (Eds.), *Handbook of volatility models and their applications*, Chapter 16. Wiley.

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: the real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*, 665–676.

Goldberger, A. and H. Theil (1961). On pure and mixed statistical estimation in economics. *International Economic Review 2*, 65–78.

Golub, G., M. Heath, and G. Wahba (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics 21*, 215–223.

Gospodinov, N. (2009). A new look at the forward premium puzzle. *Journal of Financial Econometrics 7*, 312–338.

Granger, C. (1969). Prediction with a generalized cost of error function. *Operational Research Quarterly 20*, 199–207.

Hansen, C. (2007). Asymptotic properties of a robust variance estimator for panel data when T is large. *Journal of Econometrics 141*, 597–620.

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

Hansen, P. and A. Lunde (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics 20*, 873–889.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*, 320–338.

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

Hjalmarsson, E. (2010). Predicting global stock returns. *Journal of Financial and Quantitative Analysis 45*, 49–80.

Hoerl, A. and R. Kennard (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics 12*, 55–67.

Honda, Y. (1985). Testing the error components model with non-normal disturbances. *The Review of Economic Studies 52*, 681–690.

Hsiao, C. and M. H. Pesaran (2008). Random coefficient models. In L. Mátyás and P. Sevestre (Eds.), *The Econometrics of Panel Data*, Chapter 6. Springer.

Huber, P. (1981). *Robust Statistics*. Wiley.

Hurvich, C., J. Simonoff, and C. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B 60*, 271–293.

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

Jiang, J. (1996). REML estimation: asymptotic behavior and related topics. *Annals of Statistcs 24*, 255–286.

Juhl, T. and O. Lugovskyy (2013). A test for slope homogeneity in fixed effects models. *Econometric Reviews*, forthcoming.

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

Kurtz, T. G. and P. Protter (1991). Weak Limit Theorems for Stochastic Integrals and Stochastic Differential Equations. *Annals of Probability 19*, 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*, Chapter 37.

Liew, C. (1976). Inequality constrained least-squares estimation. *Journal of the American Statistical Association 71*, 746–751.

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

Lütkepohl, H. (1996). *Handbook of Matrices*. Wiley.

Lucas, A. (1995). Unit root tests based on M estimators. *Econometric Theory 11*, 331–346.

Maynard, A. and P. Phillips (2001). Rethinking an old empirical puzzle: econometric evidence on the forward discount anomaly. *Journal of Applied Econometrics 16*, 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*, 428–457.

Mishkin, F. (1981). Are market forecasts rational? *American Economic Review 71*, 295–306.

Monteforte, L. and G. Moretti (2013). Real-time forecasts of inflation: the role of financial variables. *Journal of Forecasting 32*, 51–61.

Müller, U. and M. Watson (2008). Testing models of low-frequency variability. *Econometrica 76*, 979–1016.

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

Pesaran, M. H. and R. Smith (1995). Estimating long-run relationships from dynamic heterogenous panels. *Journal of Econometrics 68*, 79–113.

Pesaran, M. H. and T. Yamagata (2008). Testing slope homogeneity in large panels. *Journal of Econometrics 142*, 50–93.

Phillips, P. (1991). A shortcut to LAD estimator asymptotics. *Econometric Theory 7*, 450–463.

Phillips, P. (1998). New tools for understanding spurious regressions. *Econometrica 66*, 1299–1325.

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

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*, 155–159.

Poon, S. and C. Granger (2003). Forecasting volatility in financial markets: a review. *Journal of Economic Literature 41*, 478–539.

Raunig, B. (2006). The long-horizon predictability of German stock market volatility. *International Journal of Forecasting 22*, 363–372.

Shiller, R. (1973). A distributed lag estimator derived from smoothness priors. *Econometrica 41*, 775–788.

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

Swamy, P. (1970). Efficient inference in a random coefficient regression model. *Econometrica 38*, 311–323.

Toda, H. and T. Yamamoto (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics 66*, 225–250.

Toker, S., G. Şiray, and S. Kaçiranlar (2013). Inequality constrained ridge regression estimator. *Statistics & Probability Letters 83*, 2391–2398.

Ullah, A. (2004). *Finite-Sample Econometrics*. Oxford University Press.

Wand, M. (2002). Vector differential calculus in statistics. *The American Statistician 56*, 55–62.

Weiss, A. (1996). Estimating time series models using the relevant cost function. *Journal of Applied Econometrics 11*, 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*, 1455–1508.

White, H. (2001). *Asymptotic Theory for Econometricians*. Emerald.

Lebenslauf

Persönliche Daten

Geburtsdatum	11.06.1984
Geburtsort	Coesfeld
Nationalität	Deutsch

Ausbildung

Oktober 2009 - Oktober 2013	**Universität Bonn: BGSE** Promotion in Volkswirtschaftslehre
April 2004 - Oktober 2009	**Universität Bonn** Studium der Volkswirtschaftslehre Abschluss: Diplom - Volkswirt
August 2007 - Mai 2008	**University of California, Berkeley, USA** Studium der Volkswirtschaftslehre
August 1994 - Juni 2003	**Heriburg Gymnasium, Coesfeld** Allgemeine Hochschulreife

Berufliche Erfahrung

Januar 2011 - März 2012	**Universität Bonn** Wissenschaftlicher Mitarbeiter Institut für Ökonometrie und OR