Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Heavy tails of OLS

Thomas Mikosch[a], Casper G. de Vries[b,*]

[a] *Department of Mathematics, Universitetsparken 5, University of Copenhagen, Copenhagen, DK-2100, Denmark*

[b] *Department of Economics, Erasmus University Rotterdam, Rotterdam, NL-3000, The Netherlands*

ARTICLE INFO

Article history:
Available online xxxx

JEL classification:
C13
C16
C20

Keywords:
Heavy tails
OLS estimator distribution
Small sample inference

ABSTRACT

Suppose the tails of the noise distribution in a regression exhibit power law behavior. Then the distribution of the OLS regression estimator inherits this tail behavior. This is relevant for regressions involving financial data. We derive explicit finite sample expressions for the tail probabilities of the distribution of the OLS estimator. These are useful for inference. Simulations for medium sized samples reveal considerable deviations of the coefficient estimates from their true values, in line with our theoretical formulas. The formulas provide a benchmark for judging the observed highly variable cross country estimates of the expectations coefficient in yield curve regressions.

1. Motivation

Regression coefficients based on financial data often vary considerably across different samples. This observation pertains to finance models like the CAPM beta regression, the forward premium equation and the yield curve regression. In economics, macro models like the monetary model of the foreign exchange rate also yield a wide spectrum of regression coefficients.

The uncertainty in CAPM regressions was reviewed in Campbell et al. (1997, Chapter 5) and Cochrane (2001, Chapter 15). Lettau and Ludvigson (2001) explicitly model the time variation in beta. Hodrick (1987) and Lewis (1995) report wildly different estimates for the Fisher coefficient in forward premium regressions. Moreover, typical estimates of the expectation coefficient in yield curve regressions reported by Fama (1976), Mankiw and Miron (1986), and Campbell and Shiller (1991) show substantial variation over time and appear to be downward biased; Campbell et al. (1997, Chapter 10.2) provide a lucid review. The coefficient of the relative money supply in the regression of the exchange rate on the variables of the monetary model of the foreign exchange rate varies considerably around its theoretical unitary value; see for example Frenkel (1993, Chapter 4). In monetary economics parameter uncertainty is sometimes explicitly taken into account when it comes to policy decisions; see Brainard (1967) and, more recently, Sack (2000). In a random coefficient model, see Feige and Swamy (1974), the (regression) coefficients themselves are subject to randomness and therefore fluctuate about some fixed values. Estimation of the random coefficient models is reviewed in Chow (1984). Moreover, strategic decisions and rapidly changing business environments imply that one often works with a relatively short data window. Similarly, as soon as some macro variables are part of the regression, one is compelled to use low frequency data and hence a small or medium sized sample.

A possible reason for the considerable variation in estimated regression coefficients across different medium sized samples is the heavy tailed nature of the innovations distribution. It is an acknowledged empirical fact that many financial variables are much better modeled by distributions that have tails thicker than the normal distribution; see e.g. Embrechts et al. (1997, Chapter 6), Campbell et al. (1997), or Mikosch (2003) and the references therein. In small and medium sized samples the central limit theory (CLT) based standard $\sqrt{n}$-rates of convergence for the OLS parameter estimators can be a poor guide to the parameter range that may occur. But small sample results for the distribution of regression estimators are rare and exact results are difficult to obtain.

In this paper we derive *explicit* expressions for the tail of the distribution of the regression estimators in finite samples for the case that the noise distribution exhibits heavy tails.[1] Consider the simple regression model:

$$Y_t = \beta X_t + \varphi_t .$$

* Corresponding author. Tel.: +31 104088956; fax: +31 104089161.
E-mail address: cdevries@ese.eur.nl (C.G. de Vries).

[1] In large samples, given that the innovations have finite variance, CLT based results apply.

0304-4076/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2012.08.015

Suppose that the i.i.d. additive noise φ_t has a distribution with Pareto-like tails, i.e. $P(|\varphi| > s) \simeq cs^{-\alpha}$ for high quantiles s, some constant $c > 0$ and tail index $\alpha > 0$. For example, the Student-t distribution with α degrees of freedom fits this assumption. The ordinary least squares estimator of β reads

$$\widehat{\beta} = \frac{\sum_{t=1}^{n} X_t Y_t}{\sum_{t=1}^{n} X_t^2} = \beta + \rho_n, \quad \text{where } \rho_n := \frac{\sum_{t=1}^{n} \varphi_t X_t}{\sum_{t=1}^{n} X_t^2}.$$

We show that under some mild conditions, for example if the X_t are i.i.d. with a standard uniform distribution and if the φ_t follow a Student-t distribution with α degrees of freedom, then

$$P(\rho_n > x) = P(\rho_n \le -x) \sim n\, E\left[\frac{X_1}{\sum_{s=1}^{n} X_s^2}\right]^{\alpha} P(\varphi > x).$$

Note that this is a fixed *finite sample* size n cum *large deviation* x result. This relation shows that for fixed n and large x there is a strong deviation of the tail probability of ρ_n from a normal based tail which, *for fixed x and large n*, would be prescribed by the CLT. In particular, the resulting Pareto-like tails of ρ_n yield a possible explanation for the empirical observation that regression estimates often fluctuate wildly around their theoretical values.

The above Pareto-like tail probabilities can be used for statements about very high quantiles of ρ_n. Suppose $q > x$ is an even higher quantile, possibly at the border or even outside the range of the data. Then

$$q \simeq x \left(\frac{P\left(\rho_n > x\right)}{P\left(\rho_n > q\right)}\right)^{1/\alpha}.$$

Below we demonstrate that in small and medium sized samples with heavy tail distributed innovations, this approximation is considerably better than the normal (CLT) based approach. It can be used to gauge the probability of observing regression coefficients of unusual size.

The results hinge on relatively weak assumptions regarding the stochastic nature of the explanatory variable. For the linear model above we require that the joint density of the explanatory variables is bounded in some neighborhood of the origin. A restriction is the condition that the regressor be exogenous; but the regressor is not assumed to be fixed. In addition to the case of additive noise φ, we also investigate the case of random coefficients β, i.e. the case with multiplicative noise. Moreover, we allow for the possibility that the multiplicative noise component is correlated with the additive noise term. In this sense there can be correlation between the economic explanatory part and the additive noise structure. Both the noise and the regressor are allowed to be time dependent. The time dependency has an extra effect on the dispersion of the regression coefficients.

The paper does not propose alternative regression procedures such as the estimator studied in Blattberg and Sargent (1971) or the Least Absolute Deviations estimator, Rank estimators investigated in this issue by Hallin et al. (in this issue), tail trimming for GMM estimation studied by Hill and Renault (2010), the partially adaptive methods proposed in Butler et al. (1990), or the maximum likelihood procedure for the case of residuals that follow infinite variance stable distributions as considered in Nolan and Revah (in this issue); nor do we venture into the issue of model identification of infinite variance autoregressive processes as investigated in Andrews and Davis (in this issue). The purpose of our paper is different. We investigate the shape of the distribution of regression coefficients when the standard OLS procedure is applied in the case that the innovations are heavy tailed distributed. Thus while the alternative estimators are meant to overcome the deficiencies of the OLS procedure in the presence of heavy tails, we quantitatively describe the properties of the OLS procedure in this situation. The OLS method is very widely applied, including the case of financial data which are known to exhibit heavy tails. Therefore it is of interest to understand the OLS results under non-standard conditions.

The theoretical results are first illustrated by means of a simulation experiment. The Monte Carlo study demonstrates that in medium sized samples the estimated coefficients can deviate considerably from their true values. The expressions for the tail probabilities are shown to work as anticipated and their use for inference is demonstrated. Subsequently, we investigate the relevance of the theory for the wide dispersion of the expectation hypothesis coefficients in yield curve regressions.

2. The model

We study the regression model:

$$Y_t = (\beta + \varepsilon_t)X_t + \varphi_t, \tag{1}$$

where $(\varepsilon_t, \varphi_t)$ is a strictly stationary noise sequence of 2-dimensional random vectors, and (X_t) is a sequence of explanatory variables, independent of the noise. In what follows, we write φ, ε, etc., for generic elements of the strictly stationary sequences (φ_t), (ε_t), etc. The coefficient β is a fixed parameter to be estimated by regression. The model (1) comprises a large variety of different economic models since it allows for both additive and multiplicative uncertainty. If the noises ε_t and φ_t have zero mean, then, conditionally on the information at time $t - 1$, the model (1) captures the structure of many of the rational expectations finance models such as the CAPM.

In what follows, we assume that the right tail of the marginal distributions $F_\varepsilon(x)$ and $F_\varphi(x)$ of ε and φ, respectively, is regularly varying with index $\alpha > 0$. This means that the limits

$$\lim_{x\to\infty} \frac{1 - F(xs)}{1 - F(x)} = s^{-\alpha} \quad \text{for all } s > 0 \tag{2}$$

exist for $F \in \{F_\varepsilon, F_\varphi\}$. Regular variation entails that $(\alpha + \delta)$-th moments of F are infinite for $\delta > 0$, supporting the intuition on the notion of a heavy tailed distribution. Some prominent members of the class of distributions with regularly varying tails are the Student-t, F-, Fréchet, infinite variance stable and Pareto distributions. First order approximations to the tails of these distribution functions F are comparable to the tail $c\,x^{-\alpha}$ of a Pareto distribution for some $c, \alpha > 0$, i.e.,

$$\lim_{x\to\infty} \frac{1 - F(x)}{c\,x^{-\alpha}} = 1.$$

The power like decay in the right tail area implies the lack of moments higher than α. There are other distributions which have fatter tails than the normal distribution, such as the exponential or lognormal distributions. But these distributions possess all power moments. These are less suitable for capturing the very large positive and highly negative values observed in financial data sets.

Independent positive random variables $A_1, \ldots, A_n$ with regularly varying right tails (possibly with different indices) satisfy a well known additivity property of their convolutions; see for example Feller (1971). This means that

$$\lim_{x\to\infty} \frac{\sum_{i=1}^{n} P(A_i > x)}{P\left(\sum_{i=1}^{n} A_i > x\right)} = 1. \tag{3}$$

This is a useful fact when it comes to evaluating the distributional tail of (weighted) sums of random variables with regularly varying tails. The *ordinary least squares* (OLS) estimator of β is comprised of such sums, but also involves products and ratios of random variables. In particular, the OLS estimator $\widehat{\beta}$ of β in model (1) is

$$\widehat{\beta} = \frac{\sum_{t=1}^{n} X_t Y_t}{\sum_{t=1}^{n} X_t^2} = \beta + \rho_{n,\varepsilon} + \rho_{n,\varphi}, \tag{4}$$

and involves the terms

$$\rho_{n,\varepsilon} := \frac{\sum_{t=1}^{n} \varepsilon_t X_t^2}{\sum_{t=1}^{n} X_t^2} \quad \text{and} \quad \rho_{n,\varphi} := \frac{\sum_{t=1}^{n} \varphi_t X_t}{\sum_{t=1}^{n} X_t^2}. \tag{5}$$

Thus, in the case of fixed regressors X_t and with noise $(\varepsilon_t, \varphi_t)$ whose components have distributions with regularly varying tails, one can rely on the additivity property (3) to deduce the tail probabilities of the $\widehat{\beta}$ distribution. But if the regressors are stochastic, we face a more complicated problem for which we derive new results.

This paper investigates the finite sample variability of the regression coefficient estimator in models with additive noise and random coefficients when the noise comes from a heavy tailed distribution. In Section 3 we derive the finite sample tail properties of the distribution of the OLS estimator of β in model (1) when the noise has a distribution with regularly varying tails; see (2). The simulation study in Section 4 conveys the relevance of the theory. Section 5 applies the theory to the distribution of the expectations coefficient in yield curve estimation. Some proofs are relegated to the Appendix.

3. Theory

In this section we derive the finite sample tail properties of the distribution of the OLS regression coefficient estimator in the model (1) when the noise distribution has regularly varying tails. To this end we first recall in Section 3.1.1 the definitions of regular and slow variation as well as the basic scaling property for convolutions of random variables with regularly varying distributions. Subsequently, we obtain the regular variation properties for inner products of those vectors of random variables that appear in the OLS estimator of β. The joint distribution of these inner products is multivariate regularly varying. In Section 3.1.2 we give conditions for the finiteness of moments of quotients of random variables. Finally, we derive the asymptotic tail behavior of the distribution of the OLS estimator of β by combining the previous results. We present the main results on the finite sample tail behavior of $\widehat{\beta}$ for i.i.d. regularly varying noise (Section 3.2.1), for regularly varying linearly dependent noise (Section 3.2.2) and give some comments on the case of general regularly varying noise (Section 3.2.3).

3.1. Preliminaries

3.1.1. Regular variation

A positive measurable function L on $[0, \infty)$ is said to be *slowly varying* if

$$\lim_{x\to\infty} \frac{L(cx)}{L(x)} = 1 \quad \text{for all } c > 0.$$

The function $g(x) = x^{\alpha} L(x)$ for some $\alpha \in \mathbb{R}$ is then said to be *regularly varying with index* α. We say that the *random variable* X and *its distribution* F (we use the same symbol F for its distribution function) are *regularly varying with (tail) index* $\alpha \ge 0$ if there exist $p, q \ge 0$ with $p + q = 1$ and a slowly varying function L such that

$$\begin{aligned} F(-x) &= q\,x^{-\alpha}\,L(x)\,(1 + o(1)) \quad \text{and} \\ \overline{F}(x) &:= 1 - F(x) = p\,x^{-\alpha}\,L(x)\,(1 + o(1)), \quad x \to \infty. \end{aligned} \tag{6}$$

Condition (6) is usually referred to as a *tail balance condition*. For an encyclopedic treatment of regular variation, see Bingham et al. (1987).

In what follows, $a(x) \sim b(x)$ for positive functions a and b means that $a(x)/b(x) \to 1$, usually as $x \to \infty$. We start with an auxiliary result which is a slight extension of Lemma 2.1 in Davis and Resnick (1996) where this result was proved for non-negative random variables. The proof in the general case is analogous and therefore omitted.

Lemma 3.1. *Let G be a distribution function concentrated on* $(0, \infty)$ *satisfying* (6)*. Assume $Z_1, \ldots, Z_n$ are random variables such that*

$$\begin{aligned} &\lim_{x\to\infty} \frac{P(Z_i > x)}{\overline{G}(x)} = c_i^{+} \quad \textit{and} \\ &\lim_{x\to\infty} \frac{P(Z_i \le -x)}{\overline{G}(x)} = c_i^{-}, \quad i = 1, \ldots, n, \end{aligned} \tag{7}$$

for some non-negative numbers $c_i^{\pm}$ and

$$\lim_{x\to\infty} \frac{P(|Z_i| > x, |Z_j| > x)}{\overline{G}(x)} = 0, \quad i \neq j.$$

Then

$$\lim_{x\to\infty} \frac{P(Z_1 + \cdots + Z_n > x)}{\overline{G}(x)} = c_1^{+} + \cdots + c_n^{+}$$

and

$$\lim_{x\to\infty} \frac{P(Z_1 + \cdots + Z_n \le -x)}{\overline{G}(x)} = c_1^{-} + \cdots + c_n^{-}.$$

The following result is a consequence of this lemma.

Lemma 3.2. *Suppose Z_i are regularly varying random variables with tail index $\alpha_i > 0$, $i = 1, \ldots, n$. Assume that one of the following conditions holds.*

1. *The Z_i's are independent and satisfy* (7) *with* $\overline{G}(x) = P(|Z_1| > x)$, $x > 0$.
2. *The Z_i's are non-negative and independent.*
3. *Z_1 and Z_2 are regularly varying with indices $0 < \alpha_1 < \alpha_2$ and the parameters p_1, q_1 in the tail balance condition* (6) *for the distribution of Z_1 are positive.*

Then under (1) *or* (2) *the relations*

$$\begin{aligned} &P(Z_1 + \cdots + Z_n > x) \sim P(Z_1 > x) + \cdots + P(Z_n > x), \\ &P(Z_1 + \cdots + Z_n \le -x) \sim P(Z_1 \le -x) + \cdots + P(Z_n \le -x) \end{aligned}$$

hold as $x \to \infty$. If condition (3) *applies, as $x \to \infty$,*

$$\begin{aligned} &P(Z_1 + Z_2 > x) \sim P(Z_1 > x) \quad \textit{and} \\ &P(Z_1 + Z_2 \le -x) \sim P(Z_1 \le -x). \end{aligned}$$

The proof is given in the Appendix.

Recall the definition of a regularly varying random vector $\mathbf{X}$ with values in $\mathbb{R}^d$; see for example Haan et al. (1977), Resnick (1986, 1987). In what follows, $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$ with respect to a (given) norm $|\cdot|$ and $\xrightarrow{v}$ refers to *vague convergence* on the Borel σ-field of $\mathbb{S}^{d-1}$; see Resnick (1986, 1987) for details.

Definition 3.3. The random vector $\mathbf{X}$ with values in $\mathbb{R}^d$ and its distribution are said to be *regularly varying with index* α *and spectral measure* P_Θ if there exists a random vector Θ with values in $\mathbb{S}^{d-1}$ and distribution P_Θ such that the following limit exists for all $t > 0$:

$$\frac{P(|\mathbf{X}| > tx, \mathbf{X}/|\mathbf{X}| \in \cdot)}{P(|\mathbf{X}| > x)} \xrightarrow{v} t^{-\alpha} P_\Theta(\cdot), \quad x \to \infty. \tag{8}$$

The vague convergence in (8) means that

$$\frac{P\left(|\mathbf{X}| > tx, \mathbf{X}/\,|\mathbf{X}| \in S\right)}{P(|\mathbf{X}| > x)} \to t^{-\alpha} P_\Theta(S),$$

for all Borel sets $S \subset \mathbb{S}^{d-1}$ such that $P_\Theta(\partial(S)) = 0$, where $\partial(S)$ denotes the boundary of S. Alternatively, (8) is equivalent to the totality of the relations

$$\frac{P(\mathbf{X} \in xA)}{P(|\mathbf{X}| > x)} \to \mu(A).$$

Here μ is a non-null measure on the Borel σ-field of $\overline{\mathbb{R}}^d \setminus \{\mathbf{0}\}$ with property $\mu(tA) = t^{-\alpha}\mu(A)$, $t > 0$, for any Borel set $A \subset \overline{\mathbb{R}}^d \setminus \{\mathbf{0}\}$, bounded away from zero and such that $\mu(\partial(A)) = 0$.

We have the following result.

Lemma 3.4. *Assume that* $\mathbf{X} = (X_1, \ldots, X_d)'$ *is regularly varying in* $\mathbb{R}^d$ *with index* $\alpha > 0$ *and is independent of the random vector* $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ *which satisfies* $E|\mathbf{Y}|^{\alpha+\epsilon} < \infty$ *for some* $\epsilon > 0$. *Then the scalar product* $Z = \mathbf{X}'\mathbf{Y}$ *is regularly varying with index* α. *Moreover, if* $\mathbf{X}$ *has independent components, then as* $x \to \infty$,

$$\begin{aligned}
&P(Z > x) \sim P(|\mathbf{X}| > x) \\
&\quad \times \left[\sum_{i=1}^d c_i^+\, E[Y_i^\alpha I_{\{Y_i>0\}}] + \sum_{i=1}^d c_i^-\, E[|Y_i|^\alpha I_{\{Y_i<0\}}]\right], \\
&P(Z \le -x) \sim P(|\mathbf{X}| > x) \\
&\quad \times \left[\sum_{i=1}^d c_i^-\, E[Y_i^\alpha I_{\{Y_i>0\}}] + \sum_{i=1}^d c_i^+ E[|Y_i|^\alpha I_{\{Y_i<0\}}]\right],
\end{aligned} \tag{9}$$

where

$$c_i^+ = \lim_{x\to\infty} \frac{P(X_i > x)}{P(|\mathbf{X}| > x)} \quad \text{and} \quad c_i^- = \lim_{x\to\infty} \frac{P(X_i \le -x)}{P(|\mathbf{X}| > x)}.$$

The proof is given in the Appendix.

Remark 3.5. To give some intuition on Lemma 3.4, consider the case $d = 1$, i.e., $Z = X_1 Y_1$, and assume for simplicity that X_1 and Y_1 are positive random variables. Then the lemma says that

$$P(X_1 Y_1 > x) \sim EY_1^\alpha\, P(X_1 > x), \quad x \to \infty. \tag{10}$$

The latter relation is easily seen if one further specifies that $P(X_1 > x) = cx^{-\alpha}$, $x \ge c^{1/\alpha}$. Then a conditioning argument immediately yields for large x,

$$\begin{aligned}
P(Z > x) &= E[P(X_1 Y_1 > x \mid Y_1)] \\
&= \int_0^{c^{-1/\alpha}x} P(X_1 > x/y)\, dP(Y_1 \le y) + P(Y_1 > c^{-1/\alpha}x) \\
&= \int_0^{c^{-1/\alpha}x} cx^{-\alpha}\, y^\alpha\, dP(Y_1 \le y) + P(Y_1 > c^{-1/\alpha}x) \\
&= P(X_1 > x) \int_0^{c^{-1/\alpha}x} y^\alpha\, dP(Y_1 \le y) + o(P(X_1 > x)) \\
&= P(X_1 > x)\, EY_1^\alpha\, (1 + o(1)).
\end{aligned}$$

Relation (10) is usually referred to as Breiman's result; see Breiman (1965). A generalization to matrix-valued $\mathbf{Y}$ and vectors $\mathbf{X}$ can be found in Basrak et al. (2002).

3.1.2. Finiteness of moments

In this section we give conditions under which the moments of the random variables

$$\widetilde{X}_t = \frac{X_t}{\sum_{s=1}^n X_s^2}\, I_{\{X_t \ne 0\}}, \quad t = 1, \ldots, n,$$

are finite, where (X_t) is a sequence of random variables. First note that

$$|\widetilde{X}_t|^{-2} \ge I_{\{X_t \ne 0\}} \sum_{s=1}^n X_s^2 =: \widetilde{Y}_t, \quad t = 1, \ldots, n. \tag{11}$$

It will be convenient to work with the sequence $(\widetilde{Y}_t)$.

Lemma 3.6. *Let* α *be a positive number. Assume that one of the following conditions is satisfied:*

1. $X_1, \ldots, X_n$ *are i.i.d.,* $P(|X| \le x) \le cx^\gamma$ *for some* $\gamma, c > 0$ *and all* $x \le x_0$, *and* $n\gamma > \alpha$.
2. $(X_1, \ldots, X_n)$ *has a bounded density* f_n *in some neighborhood of the origin and* $n > \alpha$.

   Then $E\widetilde{Y}_t^{-\alpha/2} < \infty$ *and, hence,* $E|\widetilde{X}_t|^\alpha < \infty$ *for* $t = 1, \ldots, n$.

The proof is given in the Appendix. Condition (2) is, for example, satisfied if $(X_1, \ldots, X_n)$ is Gaussian or Student-t, and $n > \alpha$.

3.2. Tail asymptotics for regression coefficient estimators

3.2.1. i.i.d. noise

In this section we consider three sequences of random variables satisfying the following basic

Assumptions. 1. (X_t) is a sequence of random variables with $X_t \ne 0$ a.s. for every t.
2. (ε_t) is i.i.d., and ε is regularly varying with index $\alpha_\varepsilon > 0$.
3. (φ_t) is i.i.d., and φ is regularly varying with index $\alpha_\varphi > 0$.
4. (X_t) is independent of $((\varepsilon_t, \varphi_t))$.

We investigate the distributional tails of the quantities $\rho_{n,\varepsilon}$ and $\rho_{n,\varphi}$ defined in (5). Recall from (4) that the latter quantities are closely related to the OLS estimator $\widehat{\beta}$ of β in the regression model (1) with multiplicative and additive noise.

Proposition 3.7. *Assume conditions* (1), (4) *and fix* $n \ge 2$.

1. *If* (2) *holds, then as* $x \to \infty$,

$$\begin{cases}
P(\rho_{n,\varepsilon} > x) \sim P(\varepsilon > x)\, \displaystyle\sum_{i=1}^n E\left[\frac{X_i^2}{\sum_{s=1}^n X_s^2}\right]^{\alpha_\varepsilon}, \\[2ex]
P(\rho_{n,\varepsilon} \le -x) \sim P(\varepsilon \le -x)\, \displaystyle\sum_{i=1}^n E\left[\frac{X_i^2}{\sum_{s=1}^n X_s^2}\right]^{\alpha_\varepsilon}.
\end{cases} \tag{12}$$

2. *If* (3) *and, in addition,*

$$E\left(\sum_{i=1}^n X_i^2\right)^{-\frac{\alpha_\varphi}{2}-\delta} < \infty \quad \text{for some } \delta > 0, \tag{13}$$

hold, then, as $x \to \infty$,

$$P(\rho_{n,\varphi} > x) \sim P(\varphi > x) \sum_{i=1}^n E\left[\frac{X_i I_{\{X_i>0\}}}{\sum_{s=1}^n X_s^2}\right]^{\alpha_\varphi} + P(\varphi \le -x) \sum_{i=1}^n E\left[\frac{|X_i| I_{\{X_i<0\}}}{\sum_{s=1}^n X_s^2}\right]^{\alpha_\varphi},$$

$$P(\rho_{n,\varphi} \le -x) \sim P(\varphi \le -x) \sum_{i=1}^n E\left[\frac{X_i I_{\{X_i>0\}}}{\sum_{s=1}^n X_s^2}\right]^{\alpha_\varphi} + P(\varphi > x) \sum_{i=1}^n E\left[\frac{|X_i| I_{\{X_i<0\}}}{\sum_{s=1}^n X_s^2}\right]^{\alpha_\varphi}.$$

Proof. Since (X_t) and (ε_t) are independent and $X_t^2/\sum_{s=1}^n X_s^2 \le 1$, statement (1) follows from Lemma 3.4.

If (X_t) and (φ_t) are independent and $E|X_t/\sum_{s=1}^n X_s^2|^{\alpha_\varphi+\varepsilon} < \infty$ for some $\varepsilon > 0$, one can apply Lemma 3.4 to the tails of $\rho_{n,\varphi}$. An appeal to (11) and (13) ensures that this condition is satisfied, and therefore statement (2) follows. □

Sufficient conditions for condition (13) are given in Lemma 3.6. Various expressions in Proposition 3.7 can be simplified if one assumes that $X_1, \dots, X_n$ are *weakly exchangeable*, i.e., the distribution of $X_{\pi(1)}, \dots, X_{\pi(n)}$ remains unchanged for any permutation $\pi(1), \dots, \pi(n)$ of the integers $1, \dots, n$. This condition is satisfied if (X_t) is an *exchangeable sequence*. This means that (X_n) is conditionally i.i.d. If $X_1, \dots, X_n$ are weakly exchangeable, then, for example, (12) turns into

$$P(\rho_{n,\varepsilon} > x) \sim n\, P(\varepsilon > x)\, E\left[\frac{X_1^2}{\sum_{s=1}^n X_s^2}\right]^{\alpha_\varepsilon},$$

$$P(\rho_{n,\varepsilon} \le -x) \sim n\, P(\varepsilon \le -x)\, E\left[\frac{X_1^2}{\sum_{s=1}^n X_s^2}\right]^{\alpha_\varepsilon}.$$

If we assume in addition that (X_t) is stationary and ergodic, the strong law of large numbers applies to $(|X_t|^p)$ for any $p > 0$ with $E|X|^p < \infty$. This, together with a dominated convergence argument, allows one to determine the asymptotic order of the tail balance parameters in Proposition 3.7 as $n \to \infty$. We restrict ourselves to $\rho_{n,\varepsilon}$; the quantities $\rho_{n,\varphi}$ can be treated analogously. Consider

$$m_n(\alpha_\varepsilon) := \sum_{t=1}^n E\left[\frac{X_t^2}{\sum_{s=1}^n X_s^2}\right]^{\alpha_\varepsilon} = E\left[\frac{n^{-\alpha_\varepsilon}\sum_{t=1}^n |X_t|^{2\alpha_\varepsilon}}{\left(n^{-1}\sum_{s=1}^n X_s^2\right)^{\alpha_\varepsilon}}\right].$$

Assume that $E\,|X|^{2\max(1,\alpha_\varepsilon)} < \infty$. Then the law of large numbers and uniform integrability imply that as $n \to \infty$,

$$m_n(\alpha_\varepsilon) \begin{cases} \to 0 & \text{if } \alpha_\varepsilon > 1 \\ = 1 & \text{if } \alpha_\varepsilon = 1 \\ \to \infty & \text{if } \alpha_\varepsilon < 1. \end{cases}$$

We note in passing the difference between the large sample CLT based results and the fixed sample but heavy tail based result derived here. Indeed, for $\alpha_\varepsilon > 1$ (i.e., finite variance case) we have $m_n(\alpha_\varepsilon) \to 0$ as $n \to \infty$. This means that tails become thinner for large n and fixed x, whereas for n fixed and $x \to \infty$ the tails are determined by the heavy tails of ε.

Proposition 3.7 provides sufficient conditions for regular variation of $\rho_{n,\varepsilon}$ and $\rho_{n,\varphi}$. From this property we can derive our main result on the tails of the OLS estimator $\widehat{\beta} = \beta + \rho_{n,\varepsilon} + \rho_{n,\varphi}$ of the regression coefficient β.

Corollary 3.8. *If the conditions of Proposition 3.7 hold, then $\rho_{n,\varepsilon}$ and $\rho_{n,\varphi}$ are regularly varying with corresponding indices α_ε and α_φ. Moreover, if we assume that $\alpha_\varepsilon \ne \alpha_\varphi$ or that (ε_t) is independent of (φ_t) and $\alpha_\varepsilon = \alpha_\varphi$, then, as $x \to \infty$,*

$$P(\rho_{n,\varepsilon} + \rho_{n,\varphi} > x) \sim P(\rho_{n,\varepsilon} > x) + P(\rho_{n,\varphi} > x),$$
$$P(\rho_{n,\varepsilon} + \rho_{n,\varphi} \le -x) \sim P(\rho_{n,\varepsilon} \le -x) + P(\rho_{n,\varphi} \le -x),$$

where the corresponding asymptotic expressions for the tails of $\rho_{n,\varepsilon}$ and $\rho_{n,\varphi}$ are given in Proposition 3.7.

Proof. The regular variation of $\rho_{n,\varepsilon}$ and $\rho_{n,\varphi}$ is immediate from Proposition 3.7. If $\alpha_\varepsilon \ne \alpha_\varphi$, then the statement follows from the third part of Lemma 3.2. If $\alpha = \alpha_\varepsilon = \alpha_\varphi$ and (ε_t) and (φ_t) are independent, then the vector $(\varepsilon_1, \dots, \varepsilon_n, \varphi_1, \dots, \varphi_n)$ is regularly varying in $\mathbb{R}^{2n}$ with index α. Now a direct application of Lemma 3.4 yields the statement. □

Related work on the tail behavior of the OLS estimator in the linear model $\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$ can be found in He et al. (1990) and Jurečkova et al. (2001), see also the references therein. Here $\mathbf{X}$ is a design matrix, $\mathbf{b}$ the parameter vector to be estimated, $\mathbf{e}$ a vector of independent errors with common symmetric absolutely continuous distribution G. The authors consider light- and heavy-tailed cases. The heavy-tailed case is defined by assuming that the limit $\lim_{x\to\infty}(\log \overline{G}(x))/\log x^{-\alpha} = 1$ exists for some $\alpha > 0$. This condition is equivalent to $x^{-\alpha-\varepsilon} \le \overline{G}(x) \le x^{-\alpha+\varepsilon}$ for every $\varepsilon > 0$ and large $x \ge x_0(\varepsilon)$. It includes the regularly varying case. Under this condition bounds for the tails of the OLS estimator are derived. Under our slightly stronger conditions we are able to derive *explicit* expressions for these tails.

3.2.2. The noise is a linear process

In the applications we also consider sequences (ε_t) and (φ_t) of dependent random variables. We assume that (ε_t) is a linear process, i.e., there exist real coefficients ψ_j and an i.i.d. sequence (Z_t) such that ε_t has representation

$$\varepsilon_t = \sum_{j=0}^\infty \psi_j Z_{t-j} = \sum_{j=-\infty}^t \psi_{t-j} Z_j, \quad t \in \mathbb{Z}. \tag{14}$$

The best known examples of (causal) linear processes are the ARMA and FARIMA processes; see for example Brockwell and Davis (1991). Throughout we assume that $Z = Z_1$ is regularly varying with index $\alpha_Z > 0$ satisfying the tail balance condition

$$P(Z > x) = p\,\frac{L(x)}{x^{\alpha_Z}}\,(1+o(1)) \quad \text{and} \qquad P(Z \le -x) = q\,\frac{L(x)}{x^{\alpha_Z}}\,(1+o(1)), \quad x \to \infty, \tag{15}$$

for some $p, q \ge 0$ and $p + q = 1$. If the additional conditions

$$\begin{cases} \sum_{i=0}^\infty \psi_i^2 < \infty & \text{for some } \alpha_Z > 2, \\ \sum_{i=0}^\infty |\psi_i|^{\alpha_Z-\epsilon} < \infty & \text{for some } \alpha_Z \le 2, \text{ some } \varepsilon > 0, \\ EZ = 0 & \text{for } \alpha_Z > 1, \end{cases} \tag{16}$$

on the coefficients ψ_j and the distribution of Z hold, then (see Mikosch and Samorodnitsky (2000)) ε_t is regularly varying with index $\alpha_\varepsilon = \alpha_Z$ satisfying the relations

$$P(\varepsilon > x) = (1+o(1))\,P(|Z| > x) \times \sum_{j=0}^{\infty} \left[p\,(\psi_j^+)^{\alpha_Z} + q\,(\psi_j^-)^{\alpha_Z} \right], \quad x \to \infty, \tag{17}$$

and

$$P(\varepsilon \le -x) = (1+o(1))\,P(|Z| > x) \times \sum_{j=0}^{\infty} \left[q\,(\psi_j^+)^{\alpha_Z} + p\,(\psi_j^-)^{\alpha_Z} \right], \quad x \to \infty, \tag{18}$$

where x^+ and x^- denote the positive and negative parts of the real number x. This means that ε_t is regularly varying with index $\alpha_\varepsilon = \alpha_Z$, and it is not difficult to show that the finite-dimensional distributions of (ε_t) are also regularly varying with the same index α_ε.

For further discussion we also assume that φ_t is a linear process with representation

$$\varphi_t = \sum_{j=-\infty}^{t} c_{t-j}\,\gamma_j, \quad t \in \mathbb{Z}, \tag{19}$$

where (γ_t) is an i.i.d. regularly varying sequence with index $\alpha_\gamma > 0$. Assuming (16) for (c_j) instead of (ψ_j) and the tail balance condition (15) for γ instead of Z, it follows that the finite-dimensional distributions of (φ_t) are regularly varying with index $\alpha_\varphi = \alpha_\gamma$ and the relations analogous to (17) and (18) hold for the left and right tails of φ_t.

Next, we investigate the tail behavior of the quantities $\rho_{n,\varepsilon}$ and $\rho_{n,\varphi}$ under the assumption that (ε_t) and (φ_t) are regularly varying linear processes. We state our basic assumptions as follows.

Assumptions. 1. (X_t) is a sequence of random variables with $X_t \neq 0$ a.s. for every t.
2. (ε_t) is a linear process with representation (14), i.i.d. regularly varying noise (Z_t) with index $\alpha_\varepsilon > 0$ satisfying (15) and coefficients ψ_j satisfying (16).
3. (φ_t) is a linear process with representation (19), i.i.d. regularly varying noise (γ_t) with index $\alpha_\varphi > 0$ satisfying (15) (with Z_j replaced by γ_j) and coefficients c_j satisfying (16) (with ψ_j replaced by c_j).
4. (X_t) is independent of $((\varepsilon_t, \varphi_t))$.

The following result shows that $\rho_{n,\varepsilon}$ and $\rho_{n,\varphi}$ are regularly varying. Compare this result with Proposition 3.7 in the case of i.i.d. noise.

Proposition 3.9. *Assume that assumptions* (1), (4) *hold. Fix $n \ge 2$.*

1. *If assumption* (2) *holds, as $x \to \infty$, then*

$$P(\rho_{n,\varepsilon} > x)$$
$$\sim P(Z > x) \sum_{j=-\infty}^{n} E\left[\left[\frac{\sum_{t=\max(1,j)}^{n} \psi_{t-j} X_t^2}{\sum_{t=1}^{n} X_t^2}\right]^{+}\right]^{\alpha_\varepsilon} + P(Z \le -x) \sum_{j=-\infty}^{n} E\left[\left[\frac{\sum_{t=\max(1,j)}^{n} \psi_{t-j} X_t^2}{\sum_{t=1}^{n} X_t^2}\right]^{-}\right]^{\alpha_\varepsilon},$$

$$P(\rho_{n,\varepsilon} \le -x)$$
$$\sim P(Z > x) \sum_{j=-\infty}^{n} E\left[\left[\frac{\sum_{t=\max(1,j)}^{n} \psi_{t-j} X_t^2}{\sum_{t=1}^{n} X_t^2}\right]^{-}\right]^{\alpha_\varepsilon} + P(Z \le -x) \sum_{j=-\infty}^{n} E\left[\left[\frac{\sum_{t=\max(1,j)}^{n} \psi_{t-j} X_t^2}{\sum_{t=1}^{n} X_t^2}\right]^{+}\right]^{\alpha_\varepsilon}.$$

2. *If* (3) *and, in addition,* (13) *hold, then*

$$P(\rho_{n,\varphi} > x)$$
$$\sim P(\gamma > x) \sum_{j=-\infty}^{n} E\left[\left[\frac{\sum_{t=\max(1,j)}^{n} c_{t-j} X_t}{\sum_{t=1}^{n} X_t^2}\right]^{+}\right]^{\alpha_\varphi} + P(\gamma \le -x) \sum_{j=-\infty}^{n} E\left[\left[\frac{\sum_{t=\max(1,j)}^{n} c_{t-j} X_t}{\sum_{t=1}^{n} X_t^2}\right]^{-}\right]^{\alpha_\varphi},$$

$$P(\rho_{n,\varphi} \le -x)$$
$$\sim P(\gamma > x) \sum_{j=-\infty}^{n} E\left[\left[\frac{\sum_{t=\max(1,j)}^{n} c_{t-j} X_t}{\sum_{t=1}^{n} X_t^2}\right]^{-}\right]^{\alpha_\varphi} + P(\gamma \le -x) \sum_{j=-\infty}^{n} E\left[\left[\frac{\sum_{t=\max(1,j)}^{n} c_{t-j} X_t}{\sum_{t=1}^{n} X_t^2}\right]^{+}\right]^{\alpha_\varphi}.$$

The following corollary gives our main result about the tail behavior of the OLS estimator $\widehat{\beta}$ of β in the case when both (ε_t) and (γ_t) constitute linear processes.

Corollary 3.10. *If the assumptions for Proposition 3.9 hold, then $\rho_{n,\varepsilon}$ and $\rho_{n,\varphi}$ are regularly varying with corresponding indices α_ε and α_φ. Moreover, if we assume that $\alpha_\varepsilon \neq \alpha_\varphi$ or that (ε_t) is independent of (φ_t) and $\alpha_\varepsilon = \alpha_\varphi$, then, as $x \to \infty$,*

$$P(\rho_{n,\varepsilon} + \rho_{n,\varphi} > x) \sim P(\rho_{n,\varepsilon} > x) + P(\rho_{n,\varphi} > x),$$
$$P(\rho_{n,\varepsilon} + \rho_{n,\varphi} \le -x) \sim P(\rho_{n,\varepsilon} \le -x) + P(\rho_{n,\varphi} \le -x),$$

where the corresponding asymptotic expressions for the tails of $\rho_{n,\varepsilon}$ and $\rho_{n,\varphi}$ are given in Proposition 3.9.

3.2.3. More general dependent noise

In Section 3.2.2 we demonstrated how the results of Section 3.2.1 change under linear dependence. We focused on the linear process case because we were able to obtain explicit expressions for the asymptotic tail behavior of $\rho_{n,\varepsilon}$, $\rho_{n,\varphi}$ and $\widehat{\beta}$. For more complicated dependence structures, the regular variation of these quantities follows by an application of Lemma 3.4, if the finite-dimensional distributions of the noise sequences (ε_t) and (φ_t) are regularly varying.

For example, if we assume that both (ε_t) and (φ_t) constitute GARCH processes, then the finite-dimensional distributions of these processes are regularly varying with positive indices, provided some mild conditions on the noise sequences of the GARCH processes hold. We refer to Basrak et al. (2002) for the

Please cite this article in press as: Mikosch, T., de Vries, C.G., Heavy tails of OLS. Journal of Econometrics (2012), doi:10.1016/j.jeconom.2012.08.015

corresponding theory on regular variation of GARCH processes. Alternatively, one can choose (ε_t) as a GARCH process with regularly varying finite-dimensional distributions and (φ_t) as a linear process (e.g. ARMA) with regularly varying finite-dimensional distributions, or vice versa. The index of regular variation of a GARCH process is a known function of the GARCH parameters and the distribution of the noise. For GARCH processes the asymptotic behavior of the tails of $\widehat{\beta}$ cannot be given in explicit form as for linear processes; see Proposition 3.9. However, we know from Lemma 3.4 that $\widehat{\beta}$ inherits the tail index of the minimum of α_φ and α_ε.

Remark 3.11. The above single regressor theory can be readily extended to the case of multiple regressors. This approach requires the condition of multivariate regular variation as discussed in e.g. Basrak et al. (2002) and results in the multivariate analogues of the expressions from Proposition 3.7. The corresponding theory is rather technical and therefore we have chosen to omit further details.

4. Simulation study

We conduct a simulation study based on the model (1) with $\beta = 1$ to gain further insight into the theoretical results. The β is estimated by the OLS estimator $\widehat{\beta}$ from (4), with the X_t's mean corrected to allow for an intercept. First we discuss the specific cases of pure multiplicative and additive noise, then a combination of the two. Subsequently we investigate how linear and non-linear dependence in the innovations influences the dispersion of the coefficient estimates.

4.1. Pure multiplicative noise

Consider the special multiplicative case of (1)

$$Y_t = (1+\varepsilon_t)X_t$$

under the conditions:

1. (X_t) is i.i.d. $N(0, 0.04)$.
2. (ε_t) is i.i.d. $N(0, 1)$; or (ε_t) is i.i.d. Student-t distributed with $\alpha = 3$ degrees of freedom and rescaled such that the variance of the ε_t's is 1.
3. (X_t) and (ε_t) are independent.

Under these conditions Corollary 3.8 applies if ε is Student-t distributed. Indeed, as ε is regularly varying with index α, we may conclude that $\widehat{\beta} = \beta + \rho_{n,\varepsilon}$ is regularly varying with index α. Moreover, the tail asymptotics of $\rho_{n,\varepsilon}$ are described in part (1) of Proposition 3.7. Since the X_t's are i.i.d. and ε is symmetric, this means that

$$P(\rho_{n,\varepsilon} > x) = P(\rho_{n,\varepsilon} \le -x) \sim n\,E\left[\frac{X_1^2}{\sum_{s=1}^{n} X_s^2}\right]^{\alpha} P(\varepsilon > x). \tag{20}$$

We illustrate that if ε is Student distributed, the estimator $\widehat{\beta}$ has a wider dispersion in small samples than if the ε follows a Gaussian distribution with the same variance. From (20) one shows that at fixed sample size n, deep enough into the tails an increase of the degrees of freedom α (which increases the number of bounded moments), lowers the multiplicative constant $E\left[X_1^2/\sum_{s=1}^n X_s^2\right]^\alpha$ and it also reduces the tail probability $P(\varepsilon > x)$. Recall that the normal distribution is the limit distribution of the Student-t distribution as $\alpha \to \infty$. This suggests that the normal distribution is a natural benchmark against which the Student-t results can be judged.

Table 1
Slope estimates for the multiplicative model.

	Student-t			Normal		
Sample size	25	50	100	25	50	100
Mean	1.000	1.000	0.998	1.000	1.000	0.997
St. dev.	0.343	0.239	0.171	0.336	0.242	0.172
Min	−5.226	−1.514	−2.321	−0.942	0.064	0.254
0.5% quantile	−0.055	0.289	0.505	0.093	0.389	0.542
2.5% quantile	0.360	0.540	0.669	0.331	0.526	0.657
97.5% quantile	1.645	1.453	1.323	1.664	1.472	1.332
99.5% quantile	2.045	1.747	1.479	1.875	1.637	1.435
Max	5.615	4.569	2.927	2.438	2.051	1.628

We conduct 20,000 simulations of the time series $(Y_t)_{t=1,\ldots,n}$ and $(X_t)_{t=1,\ldots,n}$ for $n = 25, 50, 100$. Both the Student and normal distributed innovations ε are generated by using the same pseudo uniformly distributed random variables. Subsequently, we transform these random variables by the respective inverse distribution functions. In this way we control for sampling noise across the alternative distributions.

The results are in Table 1. The mean of the $\widehat{\beta}$ is always close to 1, the standard deviations do not vary much for fixed n across the different parameter sets. Choosing the same variance for ε under the normal and Student specifications translates into identical variance of the $\widehat{\beta}$'s across the two cases. The distribution of the estimates cross in the neighborhood of the 2.5% and 97.5% quantiles. More substantial differences between the Normal and Student based noise arise in the tails at the 0.5% and 99.5% quantiles and beyond. The Student distributed innovations generate a number of outliers in the $\widehat{\beta}$'s vis a vis the normally distributed innovations. The boxplots in Fig. 1 are quite revealing in this respect. Note that the vertical axes in the figure have different scales depending on the sample size. The table and figure also reveal the usual large sample based effect as the range of the distribution of the $\widehat{\beta}$'s shrinks with the sample size.

How to use these findings for inference about the observed uncertainty of the estimators $\widehat{\beta}$ across different samples? For example, evaluate (20) at two large quantile levels t and s. Then the ratio of the tail probabilities satisfies the relation

$$\frac{P\left(\widehat{\beta} - \beta > t\right)}{P\left(\widehat{\beta} - \beta > s\right)} \simeq \frac{P(\varepsilon > t)}{P(\varepsilon > s)} \simeq \left(\frac{t}{s}\right)^{-\alpha}, \tag{21}$$

where the second step follows from the fact that the Student distribution satisfies (6) with $L(x) = c$ for some constant $c > 0$ and $p = q = 0.5$. We continue with some elementary conclusions based on Table 1.

Consider the column Normal 25. Quite appropriately, the normal distribution at the 97.5% level with estimated standard deviation 0.336, yields the quantile $1 + 1.96 * 0.336 = 1.658$, which is very close to the recorded 1.664. Similarly, at 99.5% we find 1.875, while theoretically one should get $1 + 2.58 * 0.336 = 1.866\,9$. If instead the Student-t column was erroneously interpreted as coming from normal innovations, then at the 97.5% probability level with estimated standard deviation 0.343, one obtains the quantile $1 + 1.96 * 0.343 = 1.672$. This is still close to the recorded 1.645 quantile value. But at 99.5% we find the quantile 2.045 in the Student 25 experiment, while the normal distribution yields the quantile $1 + 2.58 * 0.343 = 1.885$. Under normality, the 2.045 quantile value is associated with the standard normal quantile $(2.045 - 1)/0.343 = 3.0466$ corresponding to 99.88%. So an excess probability of 0.5% in the Student case reduces by a factor 4 to 0.12% in the normal case. Instead, if we use the formula (21) with $t = 2.045 - 1$, $s = 1.645 - 1$, $\alpha = 3$ and $P\left(\widehat{\beta} - \beta > s\right) = 2.5\%$, we obtain

$$P\left(\widehat{\beta} - \beta > t\right) = \left(\frac{2.045 - 1}{1.645 - 1}\right)^{-3} \times 0.025 \simeq 5.878 \times 10^{-3}$$

which is close to the actual 0.995% level in the Table 1.

Please cite this article in press as: Mikosch, T., de Vries, C.G., Heavy tails of OLS. Journal of Econometrics (2012), doi:10.1016/j.jeconom.2012.08.015

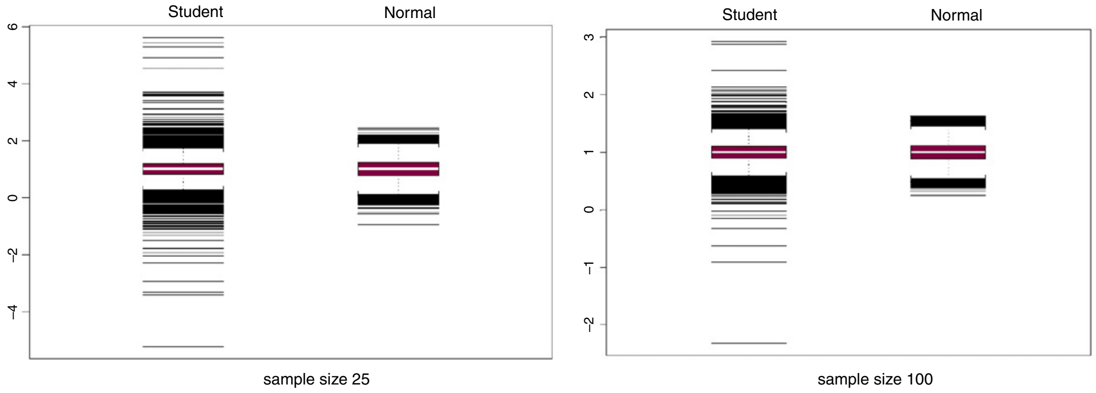


Fig. 1. Boxplots for $Y_t = (1 + \varepsilon_t)X_t$.

Furthermore, if we investigate the probability of obtaining the maximum observation 5.615, a calculation based on our formula (21) yields

$$P\left(\widehat{\beta} - \beta > t\right) = \left(\frac{5.615 - 1}{1.645 - 1}\right)^{-3} \times 0.025 \simeq 6.825 \times 10^{-5}.$$

This value is quite close to the reciprocal of the number of simulations: 0.00005. However, if the normal model were applied, the corresponding probability would be judged smaller than 10^{-40} (using the classical Laplace first order approximation based on the ratio of the density to the quantile). This approach yields an entirely different order of magnitude. Thus, if one has a number of coefficient estimates from different samples, the above formula can be used to check whether the largest and smallest estimates can reasonably be expected by using the more central outcomes and an extrapolation based on (21). This judgment can be contrasted with the usual (CLT based) opinion.

4.2. *Pure additive noise*

Next we investigate the other special case of (1), which is the specification with solely additive noise

$$Y_t = 1 + X_t + \varphi_t,$$

under the conditions

1. (X_t) is i.i.d. $N(0, 0.04)$.
2. (φ_t) is i.i.d. $N(0, 0.09)$ or (φ_t) is i.i.d. Student-t distributed with $\alpha = 3$ degrees of freedom and rescaled such that the variance of the φ_t's is 0.09.
3. (X_t) and (φ_t) are independent.

If φ is Student-t distributed, Corollary 3.10 applies. In particular, we conclude from part (2) of Proposition 3.7 that

$$P(\rho_{n,\varphi} > x) = P(\rho_{n,\varphi} \le -x) \sim n\,E\left[\frac{|X_1|}{\sum_{s=1}^{n} X_s^2}\right]^{\alpha} P(\varphi > x). \tag{22}$$

As in the case of (20), the expression (22) reveals that sufficiently deep into the tails and at fixed n, an increase in α lowers the tail probability $P(\left|\rho_{n,\varphi}\right| > x)$. We conduct again 20,000 simulations to generate the time series $(Y_t)_{t=1,\ldots,n}$ and $(X_t)_{t=1,\ldots,n}$ for $n = 25, 50, 100$. The same seeds are used as in the case of the multiplicative model to enhance comparability across the two cases. The variance of φ is chosen lower than the variance of ε of the multiplicative specification; this generates quantiles of comparable size. As before, the mean of $\widehat{\beta}$ is always close to 1, the standard deviations do not vary much for fixed n across the different parameter sets, neither are the 2.5% and 97.5% quantiles very different across the Student and normally distributed innovations. But at the more extreme quantiles we do again see from Table 2 that the heavy tail makes a difference. By the fact that the distributions of the $\widehat{\beta}$'s cross, not only does the Student noise generate more outliers, it also implies more peakedness into the center. Boxplots are provided in Fig. 2 and tell a similar story as for the case of multiplicative noise. As the different scales of the vertical axes reveal, the ranges shrink if the sample size n increases.

Table 2
Slope estimates for the additive model.

	Student-t			Normal		
Sample size	25	50	100	25	50	100
Mean	0.999	1.002	1.001	0.998	1.003	1.000
St. dev.	0.324	0.218	0.153	0.318	0.218	0.153
Min	−2.937	−1.675	−0.770	−0.406	0.123	0.323
0.5% quantile	0.024	0.361	0.587	0.146	0.415	0.611
2.5% quantile	0.372	0.578	0.706	0.363	0.569	0.700
97.5% quantile	1.621	1.419	1.300	1.626	1.423	1.299
99.5% quantile	1.996	1.611	1.429	1.859	1.570	1.393
Max	5.540	3.776	3.413	2.398	2.179	1.617

4.3. *Combination of additive and multiplicative noise*

We combine the multiplicative and additive model into the mixed model

$$Y_t = (1 + \varepsilon_t)(1 + X_t)$$

satisfying the conditions:

1. (X_t) is i.i.d. $N(0, 0.04)$.
2. (ε_t) is i.i.d. $N(0, 0.09)$ or (ε_t) is i.i.d. Student-t distributed with $\alpha = 3$ degrees of freedom and rescaled such that the variance of the ε_t's is 0.09.
3. (X_t) and (ε_t) are independent.

The mixed model has a random intercept and a random coefficient driven by the same noise ε. Hence, the tail probabilities in the case of the Student noise are the sum of (20) and (22)

$$n\left\{E\left[\frac{X_1^2}{\sum_{s=1}^{n} X_s^2}\right]^{\alpha} + E\left[\frac{|X_1|}{\sum_{s=1}^{n} X_s^2}\right]^{\alpha}\right\} P(\varepsilon > x). \tag{23}$$

The setup of the experiment is as before and we again use the same seed. The results are in Table 3 and in Fig. 3. Even though the additive noise and multiplicative noise are both present and are even perfectly correlated, the results are very similar to the two previous cases. Even if we increase the variance of ε to 1, as in the case of the pure multiplicative model, the results do not change very much (this case is not shown for consideration of space).

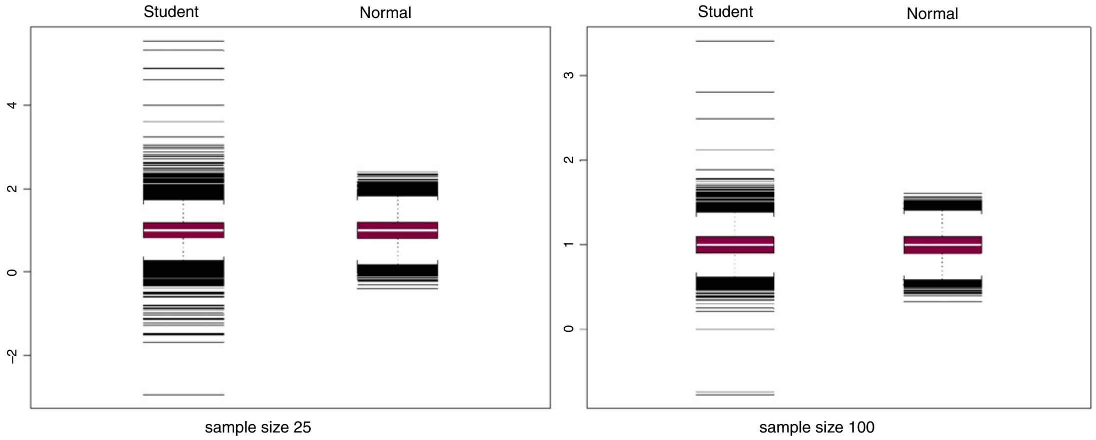


Fig. 2. Boxplots for $Y_t = 1 + X_t + \varphi_t$.

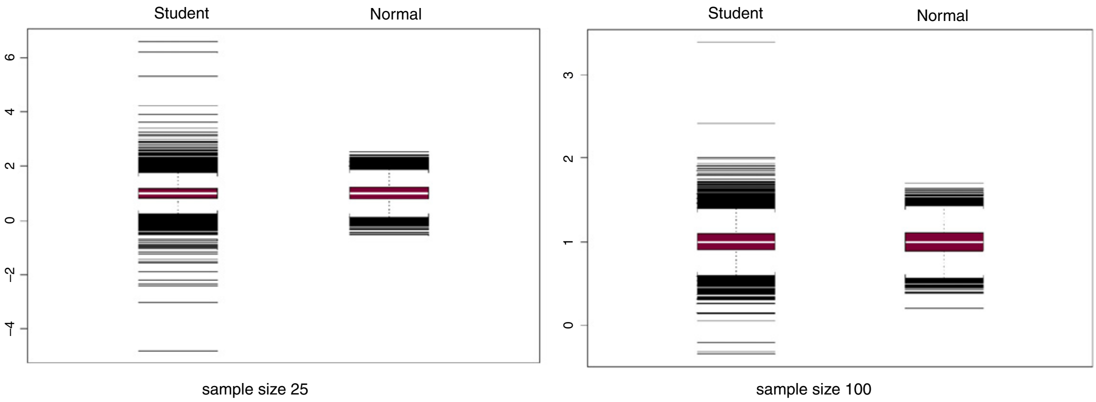


Fig. 3. Boxplots for $Y_t = (1 + \varepsilon_t)(1 + X_t)$.

Table 3
Slope estimates for the mixed model.

	Student-*t*			Normal		
Sample size	25	50	100	25	50	100
Mean	0.999	1.002	1.000	0.998	1.003	0.999
St. dev.	0.344	0.230	0.161	0.334	0.229	0.162
Min	−4.804	−2.117	−0.344	−0.535	0.054	0.201
0.5% quantile	−0.030	0.318	0.560	0.103	0.387	0.581
2.5% quantile	0.350	0.554	0.684	0.329	0.549	0.679
97.5% quantile	1.648	1.436	1.315	1.654	1.446	1.318
99.5% quantile	2.079	1.667	1.465	1.902	1.592	1.417
Max	6.603	4.611	3.390	2.538	2.280	1.697

4.4. Linearly dependent noise

We investigate the effect of autoregressive noise in the multiplicative model

$$Y_t = 1 + (1 + \varepsilon_t)X_t.$$

In what follows, we keep the assumptions 1 and 3, but we replace (2) by a particular dependence condition:

(2′) (ε_t) is an autoregressive process of order 1, i.e. AR(1), with coefficient

$$\varepsilon_t = 0.7\varepsilon_{t-1} + Z_t, \quad t \in \mathbb{Z},$$

and where (Z_t) is i.i.d. Student-*t* distributed with $\alpha = 3$ degrees of freedom and rescaled such that the variance is unity; or, alternatively, (Z_t) is i.i.d. standard Gaussian.

Under this assumption, Z_t, ε_t and $\rho_{n,\varepsilon}$ are regularly varying with index α and from the first part of Proposition 3.9 we have

$$P(\rho_{n,\varepsilon} > x) = P(\rho_{n,\varepsilon} \le -x)$$

$$\sim P(Z > x) \sum_{j=-\infty}^{n} E\left[\frac{\sum_{t=\max(1,j)}^{n} \gamma^{t-j} X_t^2}{\sum_{t=1}^{n} X_t^2}\right]^{\alpha}. \qquad (24)$$

Notice that the factor in (24) after $P(Z > x)$ is larger than in the case of i.i.d. ε_t. Therefore the heavy tails of ε have a stronger influence on the tails of $\rho_{n,\varepsilon}$. This is illustrated in Table 4 and Fig. 4. We used a different seed from the previous three cases to generate some variety; nevertheless the seed is only important for the size of the most extreme quantiles.

The autoregressive nature of the innovations ε_t makes the distribution of $\rho_{n,\varepsilon}$ more spread out than their i.i.d. counterparts, both under the Gaussian noise and the Student distributed noise. The usual large sample effect operates as before in reducing the range. To offer a somewhat different view as is conveyed by boxplots, we report in Fig. 4 histograms (the boxplots are available upon request). These histograms demonstrate nicely that not only are the coefficient estimates more dispersed, the center of the distribution is also more peaked due to heavy tailed innovations. To conclude, the time dependency in the noise increases the dispersion of the OLS estimates in samples of moderate size. But the differences between the Normal and Student noise are not very different from before. This changes when we investigate the case of non-linear dependence.

4.5. Non-linear dependency

In order to study the influence of non-linear dependence on the OLS slope estimator for the multiplicative model $Y_t = 1+(1+\varepsilon_t)X_t$

Table 4
Slope estimates for the multiplicative model with AR(1) noise.

	Student-*t*			Normal		
Sample size	25	50	100	25	50	100
Mean	1.001	1.002	1.000	1.007	1.003	1.003
St. dev.	0.719	0.529	0.383	0.718	0.532	0.380
Min	−14.257	−5.659	−3.361	−2.109	−1.205	−0.332
0.5% quantile	−1.050	−0.483	−0.041	−0.838	−0.403	0.026
2.5% quantile	−0.348	−0.020	0.255	−0.399	−0.052	0.253
97.5% quantile	2.375	2.025	1.719	2.414	2.035	1.741
99.5% quantile	2.988	2.472	2.031	2.867	2.378	1.976
Max	8.170	7.193	6.944	4.126	3.339	2.406

Note: Monte Carlo results for $Y_t = 1 + (1 + \varepsilon_t)X_t$, with $\varepsilon_t = 0.7\varepsilon_{t-1} + Z_t$, and where Z is either standard normal or unit variance Student-*t* with 3 d.f. distributed.

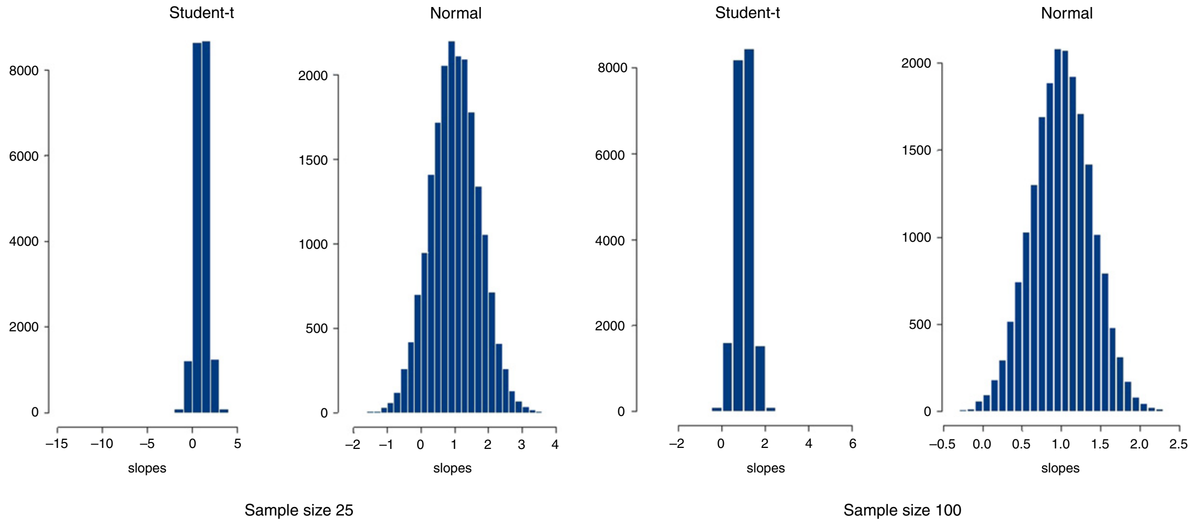


Fig. 4. Histograms for $Y_t = 1 + (1 + \varepsilon_t)X_t$, with $\varepsilon_t = 0.7\varepsilon_{t-1} + Z_t$.

Table 5
Slope estimates with ARCH(1) noise.

Model	Multiplicative noise				Additive noise			
	Student-*t*		Normal		Student-*t*		Normal	
Sample size	25	100	25	100	25	100	25	100
Mean	1.000	1.000	1.000	1.000	0.998	1.000	0.999	1.000
St. dev.	0.150	0.079	0.100	0.051	0.478	0.236	0.318	0.152
Min	−0.555	−0.589	0.515	0.766	−5.383	−1.417	−0.333	0.334
0.5% quantile	0.482	0.753	0.734	0.866	−0.636	0.281	0.142	0.601
2.5% quantile	0.711	0.851	0.802	0.898	0.063	0.548	0.363	0.703
97.5% quantile	1.285	1.150	1.197	1.102	1.930	1.452	1.630	1.301
99.5% quantile	1.490	1.240	1.261	1.135	2.573	1.714	1.839	1.392
Max	3.315	1.877	1.561	1.238	7.797	4.209	2.646	1.557

Note: Monte Carlo results for $Y_t = 1 + (1 + \varepsilon_t)X_t$ and $Y_t = 1 + X_t + \varphi_t$ with ARCH(1) innovations in ε and φ.

and the additive model $Y_t = 1 + X_t + \varphi_t$, we consider an ARCH(1) process in the innovations ε *and* φ respectively. The two ARCH(1) processes read

$$\varepsilon_t = \sigma_t Z_t, \qquad \sigma_t^2 = \beta + \lambda\, \varepsilon_{t-1}^2,$$

and

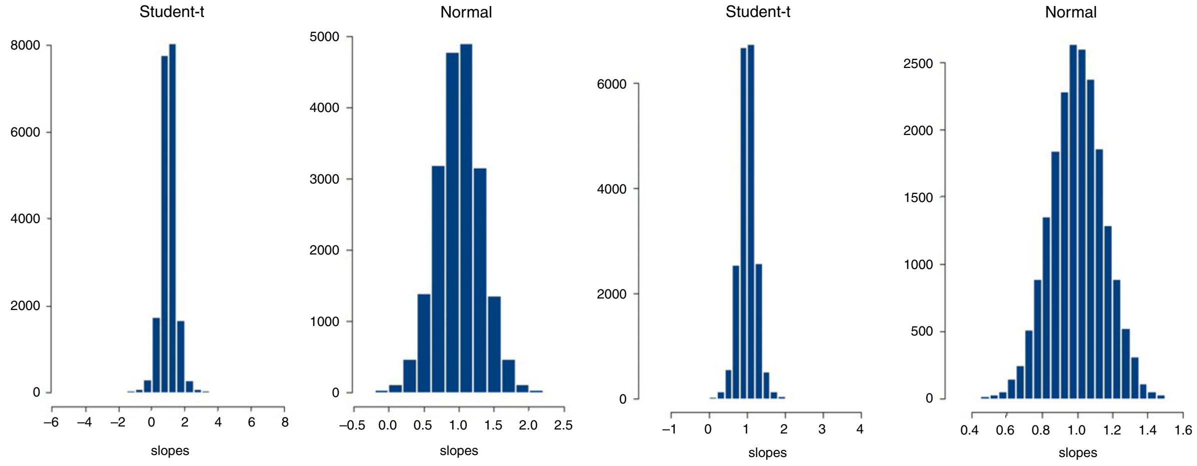
$$\varphi_t = \sigma_t Z_t, \qquad \sigma_t^2 = \beta + \lambda\, \varphi_{t-1}^2,$$

for i.i.d. $N(0, 1)$ noise (Z_t). It follows from e.g. Basrak et al. (2002) that ε and φ are regularly varying with index α which is given as the unique solution to the equation

$$E|\lambda Z^2|^{\alpha/2} = 1.$$

For $\alpha = 3$, this boils down to solving $\Gamma(2) = \sqrt{\pi}\,(2\lambda)^{-3/2}$, which gives $\lambda \simeq 0.732$, see e.g. Embrechts et al. (1997, Chapter 8). Setting $\beta = 0.09\left(1 - \frac{1}{2}\,(\pi)^{1/3}\right) \simeq 0.024$, then induces an unconditional variance $\beta/(1 - \lambda)$ of 0.09. The simulation setup is as before, except that in order to initialize the ARCH process, the first 5 realizations are ignored. For comparison we also draw an i.i.d. $N(0, 0.09)$ sample of φ_t's and ε_t's. Independently of (φ_t) and (ε_t) we drew i.i.d. $N(0, 0.04)$ random variables X_t. Subsequently we calculate the 20,000 values of the OLS coefficient estimates.

The results are in Table 5. The differences between the heavy tailed cases with ARCH additive noise and the cases of i.i.d. normal additive noise are more pronounced than the cases with multiplicative noise. The histograms for the additive model are reported in Fig. 5. The differences for the additive case are also more pronounced in comparison with the previous cases generated by i.i.d. and linear noise. The distributions are much more spread out in the case of ARCH noise in comparison the normal i.i.d. noise. To conclude, in the case of additive noise case, the OLS coefficient estimates vary more in medium sized samples if the noise process is of the ARCH variety.



Sample size 25

Sample size 100

Fig. 5. Histograms for$Y_t = 1 + X_t + \varphi_t$ with ARCH(1) innovations in φ.

5. Application

The motivation for this study comes from the empirical fact that in various applications involving small and medium size samples of financial data there appears a wide variation in reported coefficient estimates across different samples. One possible explanation for this variation is the regular variation of the distributions of the innovations. Both from an estimation point of view and for policy making it is important to capture the uncertainty in the estimates, as standard central limit type of error bands can be quite misleading in smaller samples.

The application focuses on the yield curve. Mankiw and Miron (1986) report typical slope estimates of the expectations coefficient in yield curve regressions for quite different samples. There are two important data features. First, all reported point estimates are less than two, and two slope estimates are even negative (one is significantly negative). Only in one of their five samples, with about 80 observations each, the point estimates come close to the benchmark theoretical slope of two (zero term premium). Second, the point estimates come in a wide range. Fama and Bliss (1987) emphasize the variability of the coefficient estimates. While a sizable literature has focused on the apparent downward bias, in this study we focus on the coefficient variability.

5.1. Economics

To introduce the issue of yield curve estimation, consider a stylized three period investment problem under uncertainty:

$$\max EU(x, X_i) = \pi V(x) + (1-\pi)\Sigma_i \pi_i \rho V(X_i), \quad \Sigma_i \pi_i = 1 \qquad (25)$$

subject to $w = d + b$,

$x = (1+r)d + (1+q)b$,

$X_i = (1+s)b + (1+R_i)(1+r)d, \quad i = 1, \dots, n.$

Here $V(\cdot)$ and $\rho V(\cdot)$ are the strictly concave first and second period utility functions and where $\rho \leq 1$ is the pure rate of time preference. There are two types of uncertainty. The first type of uncertainty stems from the uncertain liquidity needs of agents as in Diamond and Dybvig (1983): With probability π an agent wants to consume early, and with probability $1 - \pi$ an agent finds out that he desires to consume late during the third period.

The other uncertainty, π_i, pertains to the return R_i on the second period short term bond. At the beginning of the first period, wealth w can be invested in a one-period bond d yielding $1 + r$ in the second period and a two-period zero-coupon bond b yielding $1+s$ in the third period. The one-period bond investment can be rolled over into a new one-period bond investment if one turns out to be a late consumer. The interest rate on the second period short term bond is uncertain at the time when the first investment decision has to be taken; R_i materializes with probability π_i, and there are n states of the world $i = 1, \dots, n$. If the consumer has early liquidity needs, the long term bond investment has to be liquidated early. This comes at a cost, and we assume that $q < r$, where $r - q$ is the liquidation premium. The costs of early liquidation are due to irreversibilities of longer term capital investments and transactions costs, see Diamond and Dybvig (1983). If $(1 + s) > (1 + r)E[1 + R_i]$, a risk averse agent will want to hedge against liquidity needs by partly investing in the long term bond and partly investing in the short term bond.

The first order conditions for a maximum to the problem (25) imply the following pricing kernel

$$E[M_i(1+R_i)] = \frac{q-r}{1+r} + E\left[M_i \frac{1+s}{1+r}\right], \qquad (26)$$

where

$$M_i = \rho \frac{1-\pi}{\pi} \frac{\partial V(X_i)/\partial X_i}{\partial V(x)/\partial x}$$

is the intertemporal marginal rate of substitution or discount factor, for short.

Rewrite (26) by using the covariance $\mathrm{cov}(M_i, R_i - r)$ between the discount factor and the short term interest rate innovations. Let y be the 1-period yield on the two-period bond, i.e., $(1+y)^2 = (1+s)$. Then (26) can be restated as

$$E[R_i - r] = 2(y - r) + P(M_i, R_i, q) + T(y, r), \qquad (27)$$

and where the term premium P and convexity term T read respectively

$$P(M_i, R_i, q) = \frac{1}{EM_i}\frac{q-r}{1+r} + \frac{\mathrm{cov}(-M_i, R_i - r)}{EM_i},$$

$$T(y, r) = \frac{(y-r)^2}{1+r}.$$

Thus the expected difference between two subsequent short rates is equal to twice the difference between the long yield and the short rate plus the term premium P and a small convexity term T. The term premium consists of two parts, a liquidity premium and a risk premium. The liquidity premium is negative, since $q < r$, which reflects the costs of liquidating the long term investment early. If agents have power utility preferences and are risk averse,

Please cite this article in press as: Mikosch, T., de Vries, C.G., Heavy tails of OLS. Journal of Econometrics (2012), doi:10.1016/j.jeconom.2012.08.015

Table 6
Slope estimates for expectations hypothesis model.

Count	OLS	Relative	Relative & convex	Pool & relative	Pool & relative & convex
Bel	1.01	1.86	0.58	0.99	0.54
Can	1.20	1.91	2.14	1.98	2.09
Ger	1.28	0.97	0.22	1.00	0.47
Den	2.02	2.21	2.27	1.44	1.93
Fr	1.99	2.89	2.96	2.68	3.35
UK	0.87	1.37	1.35	1.15	1.28
It	0.98	1.30	1.27	1.22	0.97
Jap	0.95	1.02	1.26	1.15	1.28
Aus	0.51	1.39	1.40	1.25	1.46
Por	0.54	0.70	0.42	0.42	0.49
Swe	0.84	1.29	1.16	1.05	1.10
USA	0.36	0.91	1.10	0.66	1.09
Swit	1.15	1.01	0.99	1.00	0.91
Neth	1.22	/	/	/	/
Mean	1.06	1.44	1.31	1.23	1.30
Range	1.66	2.19	2.74	2.26	2.88
St. dev.	0.48	0.61	0.76	0.57	0.79

Note: The table records coefficient estimates for the slope coefficient of the expectations model. The first column are the per country OLS results for (28); column two is for (29) without the convexity term and the third column is for (29). The base country in the regressions is The Netherlands.

then the risk premium part is positive.[2] Thus the term premium P can be of either sign, since the liquidity premium is negative and the risk premium is positive.

Not much can be said, however, about the correlation between P and the yield differential $y - r$. As the term premium is unobserved, this may cause an omitted variable bias in regressions of $R_i - r$ on $y - r$. This is the standard explanation for the "downward bias" in the regression

$$R_i - r = \theta + \beta(y - r) + \varepsilon, \qquad E\varepsilon = 0. \tag{28}$$

Typically the hypothesis $\beta = 2$, known as the *Expectations Hypothesis*, applies if agents are risk neutral and when there is no premium to liquidity in (27), is rejected as the estimated coefficients are usually significantly below 2. In this study we take $\beta < 2$ as a stylized fact. Instead, we focus on the considerable dispersion of the reported β estimates in different samples.

In the application, we first run the specification (28) for a number of different countries, with and without the convexity term, and basically find the same results as are reported in the literature. Since the model applies to each country individually, one can also investigate the model relative to a benchmark country. Due to a negative correlation between the unobserved term premium and the spread, the relative specification may diminish the omitted variable bias. The bias would be reduced if $\text{cov}(X, P) + \text{cov}(X^*, P^*)$ is negative and $-\text{cov}(X, P^*) - \text{cov}(X^*, P)$ is positive, which happens when countries experience simultaneously similar movements in their yield curves, so that P (and P^*) also co-vary negatively with $y^* - r^*$ and $y - r$ respectively. Thus we also estimate

$$\widetilde{R}_i - \widetilde{r} = \theta + \beta(\widetilde{y} - \widetilde{r}) + \tau \widetilde{T}(y, r) + \widetilde{\varepsilon}, \tag{29}$$

where $\tilde{X} = X - X^*$, and X^* denotes the benchmark country variable.

5.2. Time series regressions

We obtained data on the one month and two month interest rates from February 1995 to December 1999 for 14 countries, yielding 59 observations per country. The countries with the abbreviations used in the tables are respectively Belgium (Bel), Canada (Can), Germany (Ger), Denmark (Den), France (Fr), United Kingdom (UK), Italy (It), Japan (Jap), Austria (Aus), Portugal (Por), Sweden (Swe), United States of America (USA), Switzerland (Swit), and The Netherlands (Neth). The data are beginning of the month figures on short term treasury paper available from datastream. Since for the purpose of the paper we are interested in running both time series, panel and cross section regressions, we are squeezed by data availability. Only since the middle of the 1990s there are more than 10 countries which report such rates. By the end of that decade, though, we are confronted by the data squeeze due to the European monetary unification process, which implies that at the short end of the yield curve rates became about equal.

The first column of Table 6 gives the results for specification (28). The next column labeled relative is based on (29) but without the convexity term, and where The Netherlands is taken as the base country. The third column labeled relative & convex is based on the full specification of (29) including the convexity term $T(y, r)$. The last two columns repeat the same exercise but use panel regressions in which the data are pooled. The table conveys two main results. (i) The β-coefficients are almost all significantly positive, but hover more closely around 1, rather than around the Expectations Hypothesis value 2, indicating that the term premium $P(M_i, R_i, q)$ is negatively correlated with $(y - r)$. Comparing the first column of Table 6 to the other columns, it appears that using the relative country spreads is helpful in reducing the downward bias. (ii) There is quite a considerable variation in the coefficient estimates as can be seen from the range statistic, which is the difference between the highest and the lowest estimated value. These results thus corroborate the results reported previously in the literature, although in our sample we do not find any negative coefficients. Not reported are the intercept estimates which were invariably very small and never significantly different from zero.

5.3. Cross section regressions and coefficient variability

We investigate further the considerable variation in the individual cross country estimates. Fama and Bliss (1987) and Cochrane (2001, Chapter 20) show that almost all variation is in the short term interest innovations $R_i - r$ rather than in the variation of the expected changes, i.e. $\sigma_{R_i - r} \gg \sigma_{y-r}$. Could it be that the news distribution is heavy tailed in such a way that it explains the considerable variation in the slope estimates?

Since several countries pursued quite similar economic (monetary) policies during the second half of the nineties, the variation is perhaps not so much a cross country issue. Rather, the variation may be due to similar country shocks over time. To investigate the amount of time variation in the slopes, we now assume

[2] This can be shown by using that for any random variable $X \geq 0$ and a monotone non-decreasing function $g(x)$ it follows from association that the relation $E[X\,g(X)] \geq E[X]\;E[g(X)]$ holds; see Resnick (1987, Lemma 5.32(iv)).

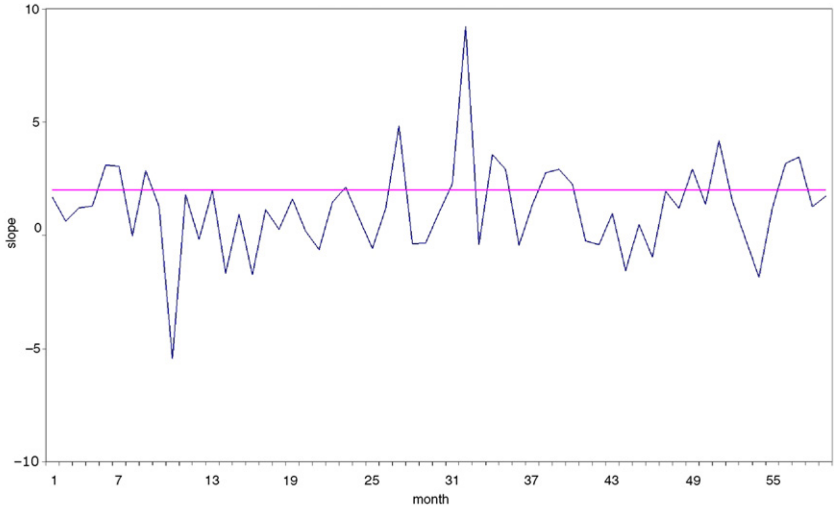


Fig. 6. Cross section slope estimates.

Table 7
Summary statistics cross country slope estimates.

Mean	1.18
Standard error (of mean)	0.26
Median	1.21
Standard deviation	2.03
Sample variance	4.13
Kurtosis	4.52
Skewness	0.50
Range	14.65
Minimum	−5.43
Maximum	9.22
Mean R^2	0.31
Standard error (of mean R^2)	0.03
Standard deviation R^2	0.22

Note: The table records the summary statistics of the cross country regressions.

identical slopes per country but allow for temporary disturbances. This model can be estimated by running per period cross section regressions. A graph of the 59 cross country $\widehat{\beta}_t$ OLS estimates for the specification (29) is provided in Fig. 6. There is indeed quite a bit of variation in the cross section slope estimates. Since the yield curve is widely known to be rotating and shifting, this was perhaps to be expected. Summary statistics are given in Table 7. The standard error of the slope estimates, the range and the kurtosis confirm the sizable variation. But this is not due to a bad fit, since the R-squared statistic is mostly acceptable for the small cross section regressions. The variation appears to be genuine. The average of the slope estimates re-confirms the downward bias from the Expectations Hypothesis.

The high kurtosis reported in Table 7 points towards the possibility of a heavy tailed distribution. In Section 3 we showed that the heavy tail feature of the innovations carries over to the distribution of the coefficient estimates. To pursue this further, we investigate the tail shape of the empirical distribution of the 59 cross section $\widehat{\beta}_t$'s.

To be able to exploit (20) or (22), as was done in (21), one first needs an estimate of the coefficient of regular variation α. The standard approach to estimating α is by means of computing the Hill statistic, which coincides with the maximum likelihood estimator of the tail index in case the data are exactly Pareto distributed. If the Pareto approximation is only good in the tail area, one conditions the estimator on a high threshold s, say, to obtain

$$\widehat{\frac{1}{\alpha}} = \frac{1}{M} \sum_{i=1, X_i < -s} \log \frac{-X_i}{s}, \tag{30}$$

where M is the random number of extreme observations X_i that fall below the threshold $-s$. In practice one of the higher order statistics is used as a threshold. If one plots $\widehat{\alpha}$ derived from (30) against different threshold levels, one obtains the so called Hill plot. Since the Hill estimator is biased, stemming from the fact that the distribution is not exactly Pareto, there is a region where if one uses too many observations the bias dominates, while the variance exerts a dominating influence if one uses too few observations. There exists an intermediate region in which the two vices are optimally balanced. The way to read these plots is further explained in Embrechts et al. (1997, p. 341).

The 59 slope estimates depicted in Fig. 6 are the basis for the Hill plot. A Hill plot gives the value of the inverse of (30) along the vertical axis, and the horizontal axis reports the number of highest order statistics used in (30). First, however, the slopes are demeaned by their empirical mean of 1.18. Secondly, we use the absolute values of the demeaned slopes upon the presumption that the slope distribution is symmetric. This helps to increase the low number of observations. The Hill plot in Fig. 7 shows that there is quite some variation if one uses only the most extreme observations, while moving to the right the bias kicks in and causes the monotone decline. Eyeballing this plot suggest that $\alpha \simeq 2$.

Given the low number of observations, we also use the regression methodology to estimate α as a robustness check. To this end Fig. 8 offers a Pareto log–log plot of the cross country slope estimates. In a Pareto log–log plot, the log rank of the data is plotted against the log of the slope estimates. A regression on the most rightward (largest slopes) 26 observations gives a slope estimate of −1.81, a standard error of 0.09, and R^2 of 0.95. This corresponds to an $\alpha = 1.81$.

If we proceed with $\alpha = 1.8$, we find that if we use as a benchmark the 10% largest observation (slope value 2.92) that at the 5% level one should find

$$t \simeq 2.92 \times \left(\frac{0.1}{0.05}\right)^{1/1.8} \simeq 4.29.$$

The demeaned 5% highest slope estimate is in fact somewhat lower 3.63. At the 1.5% probability level (corresponding to 1/59), which is the highest slope estimate, we get a demeaned slope of 8.02, while the approximation formula suggests

$$t \simeq 2.92 \times \left(\frac{0.1}{0.015}\right)^{1/1.8} \simeq 8.38.$$

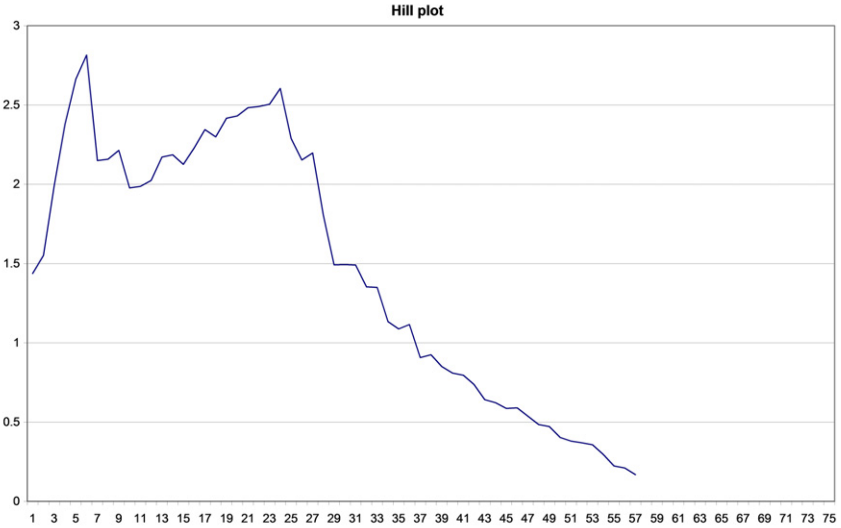


Fig. 7. Hill plot cross section slope estimates.

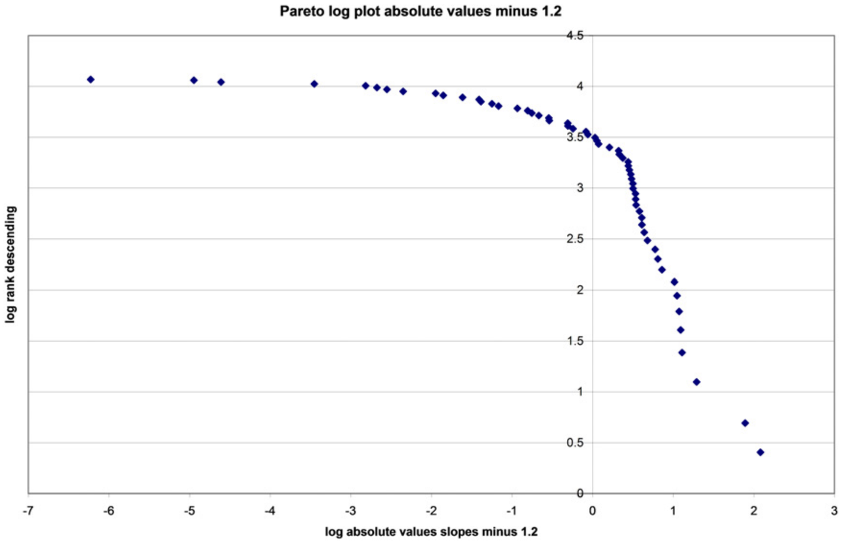


Fig. 8. Pareto log–log plot of the cross section slope estimates.

The corresponding normal based calculation would give

$$P\left(\widehat{\beta} - \beta > 8.02\right) \sim \frac{\sigma}{x}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right)$$
$$= \frac{2.03}{(8.02)}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{8.02}{2.03}\right)^2\right)$$
$$\simeq 4.\,12 \times 10^{-5}$$

which is much lower than 1.5×10^{-2}. So from the fat tail point of view, the variation and extremes in the slope estimates are quite normal, but not so under the normal interpretation.[3]

Our theory can also be used to test the null Expectations Hypothesis that $\beta = 2$. Assume that the cross section coefficient estimates depicted in Fig. 6 and summarized in Table 7 are i.i.d. This assumption is reasonable since, as discussed below, the data reveal strong cross sectional but little intertemporal dependence. The average of the estimates is 1.18. To be able to use our tail approximations, we have to take care of the change in the scale parameter due to averaging. Assuming that $P(\widehat{\beta}_i > t) \simeq ct^{-\alpha}$, a standard result on convolutions (see e.g. Lemma 1.3.1 in Embrechts et al. (1997)) yields

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\beta}_i > t\right) \simeq cn^{1-\alpha}t^{-\alpha}.$$

Proceeding as above, we find $t \simeq 6.31$. Rescaling according to the above formula then gives $6.31 \times 0.037 = 0.23$. The upper bound of a two-sided confidence band at the 5% level is $1.18 + 0.23 = 1.41$, which is far below the value $\beta = 2$. A similar calculation yields an upper bound of the two-sided confidence band at the 1% level of $1.18 + 0.57 = 1.75$, which is still not in agreement with the Expectations Hypothesis $\beta = 2$.

We investigate the dependency structure of the data, both over time and cross sectionally. From the correlograms for the slope estimates and the squares thereof, there is no indication for the level or the squares to be time dependent. We further investigate the autocorrelation structure by estimating an AR(1) model for the level and the squares, but find no indication that these slope estimates are time dependent. Secondly, we look at the time

[3] The Pareto plot in Fig. 8 is approximately linear with slope estimate 1.97 over the range of the 20% largest observations. Thus if we alternatively used these 20% largest observations as a benchmark, our tail approximation formula yields $t \simeq 2.90$ at the 5% level, while $t \simeq 5.65$ at the 1.5% probability level.

structure of the per country residuals from the cross sectional slope estimates, to see whether there is evidence for (time) dependency in the noise. On basis of correlograms, ARMA and ARCH estimates, there is again little or no evidence for a time structure in the residuals. To conclude, it appears that there is little or no evidence for dependency over time in the data.

Lastly, we also investigate the per period cross-sectional dependency. Since several countries pursued similar monetary policies at the time (relative to the base country), cross-sectional dependency would be consistent with the economic facts. From the cross-section slope estimates we construct the matrix (across months and countries) of residuals and the corresponding covariance matrix is formed. Using the Anderson (1984, p. 434–437) χ^2-based test, the null of cross sectional independence is overwhelmingly rejected.[4] To conclude, while there is little or no dependency over time, there is clear evidence for cross-sectional dependency. This cross sectional dependence makes that large shocks, which stem from the non-normal error structure, appear simultaneously in several countries. As is explained in the theory section and the simulation study, this may be the cause for the high variability in the small sample cross section slope estimates as depicted in Fig. 6.

6. Conclusion

The paper provides a theory for tail probabilities for the linear regression estimator distribution in medium sized samples if the multiplicative and additive error terms follow a heavy tailed distribution. We show that even if standard moment conditions such as the existence of the variance are satisfied, the usual benchmarks based on central limit theory can be misleading. The results hinge on relatively weak assumptions regarding the stochastic nature of the explanatory variable. With additive uncertainty we require that the joint density of the explanatory variables is bounded in some neighborhood of the origin. A restriction is the condition that the regressor be exogenous. On the other hand, we allow for the possibility that the random multiplicative noise component be correlated with the additive noise term, and in this sense there can be correlation between the economic explanatory part and the additive noise structure. Moreover, both the noise and the regressor are allowed to be time dependent.

It is shown that for a fixed sample size and if the perturbations are regularly varying, the OLS regression coefficient estimator has a tail probability which is the product of the tail probability of the perturbations and the expected ‘kernel weight’. From this result we obtain explicit expressions for the tail probabilities of the distribution of the OLS estimator. These formulas are useful for inference. In large samples the tail influence of the perturbation term is lost by virtue of the central limit theorem. To derive these results we started with a novel result on scaling properties of products and ratios of regularly varying random variables.

A Monte Carlo study showed the importance of the alternative assumptions of normally distributed versus heavy tail distributed innovations. Regardless of whether the noise in the regression is additive or multiplicative, there exists a clearly discernible effect of wider spread of the OLS estimator in medium sized samples, in contrast to the normal approximation and in contrast to normally distributed noise. Dependency gives an additional kick. The considerable deviations of the coefficient estimates from their true values correspond well with our theoretical formulas.

The application to yield curve estimation demonstrates the relevance of the theoretical results. Traditional slope estimates are highly variable and clearly downward biased in comparison with the theoretical value suggested by the Expectations Hypothesis, but are not necessarily in conflict with more elaborate economic theory that includes a liquidity premium. We focused on the considerable variation over time of the coefficient estimates. A thorough analysis of the set of cross country slope estimates revealed that the estimated random components come from a distribution with a regular varying tail. The heavy tail feature is significant in the cross sectional estimates due to the small cross section sample size and the cross sectional dependence. The more extreme estimates seem to adhere well to our theoretical formulas for the distribution of tail realizations.

We conclude that the theory seems applicable to economic data and potentially explains the wide variability of observed regression estimates.

Acknowledgments

Thomas Mikosch’s research is partly supported by the Danish Natural Science Research Council (FNU) Grants 09-072331 “Point process modelling and statistical inference” and 10-084172 “Heavy tail phenomena: Modeling and estimation”. Both authors are grateful to the stimulating research environments provided at EURANDOM in Eindhoven (The Netherlands) and at the Newton Institute in Cambridge (UK) in July 2001 during the thematic period on Managing Risk and Uncertainty, where early parts of this paper were written. We like to thank Martijn van Harten and Olaf van Veen for their valuable research assistance; the comments and suggestions by the referees, associate editor and Catalin Starica were very stimulating for the revision of the paper.

Appendix

We give the proofs to Lemmas 3.2, 3.4 and 3.6.

A.1. *Proofs to Section 3*

We start with Lemma 3.2.

Proof. Assume first that the Z_i’s are independent. If the Z_i’s are non-negative the result is standard; see Feller (1971, p. 278), or Embrechts et al. (1997, Lemma 1.3.1 and Appendix A3.3). For general Z_i and Z_j, $i \neq j$, using the independence,

$$\lim_{x\to\infty} \frac{P(|Z_i| > x, |Z_j| > x)}{\overline{G}(x)} = 0.$$

Taking into account these calculations and assumption (7), we see that the conditions of Lemma 3.1 are satisfied and so it follows that

$$\lim_{x\to\infty} \frac{P(Z_1 + \cdots + Z_n > x)}{\overline{G}(x)} = c_1^+ + \cdots + c_n^+,$$

implying that

$$\begin{aligned} P(Z_1 + \cdots + Z_n > x) &= (1 + o(1))\,(P(Z_1 > x) \\ &\quad + \cdots + P(Z_n > x)). \end{aligned}$$

The case of the left tail $P(Z_1 + \cdots + Z_n \leq -x)$ is analogous.

For dependent Z_1, Z_2 with $\alpha_1 < \alpha_2$, $P(|Z_2| > x) = o(P(|Z_1| > x))$. Then similar calculations as above yield that the assumptions of Lemma 3.1 are satisfied. □

We continue with Lemma 3.4.

[4] The estimation effects are neglected.

Proof. The fact that Z is regularly varying with index α follows by a straightforward application of Proposition A.1 in Basrak et al. (2002). The techniques used there also allow one to derive (9), which relation we show in detail. Let $M > 0$. Then

$$\begin{aligned} P(Z > x) = &\ P(Z > x, A_1) + P(Z > x, A_2) + P(Z > x, A_3) \\ =: &\ p_1 + p_2 + p_3, \end{aligned}$$

where

$$A_1 = \left\{|\mathbf{Y}| \le M^{-1}\right\}, \qquad A_2 = \left\{M^{-1} < |\mathbf{Y}| \le M\right\},$$
$$A_3 = \{|\mathbf{Y}| > M\}.$$

Then we have

$$p_1 \le P(|\mathbf{X}|\,|\mathbf{Y}| > x, A_1) \le P(|\mathbf{X}| > xM).$$

Clearly, $|\mathbf{X}|$ is regularly varying with index α, and so

$$\lim_{M\to\infty} \limsup_{x\to\infty} \frac{p_1}{P(|\mathbf{X}| > x)} = 0. \tag{31}$$

Recall from Breiman (1965) that for independent non-negative random variables ξ, η such that $E\eta^{\alpha+\delta} < \infty$ for some $\delta > 0$ and $P(\xi > x)$ regularly varying with index $\alpha > 0$.

$$P(\xi\eta > x) \sim E\eta^\alpha P(\xi > x), \quad x \to \infty.$$

For p_3 we have, using Breiman's result, regular variation of $|\mathbf{X}|$ and Lebesgue dominated convergence,

$$\begin{aligned} &\lim_{M\to\infty} \limsup_{x\to\infty} \frac{p_3}{P(|\mathbf{X}| > x)} \\ &\le \lim_{M\to\infty} \limsup_{x\to\infty} \frac{P(|\mathbf{X}|\,|\mathbf{Y}|\, I_{(M,\infty)}(|\mathbf{Y}|) > x)}{P(|\mathbf{X}| > x)} \\ &= \lim_{M\to\infty} E[|\mathbf{Y}|^\alpha\, I_{(M,\infty)}(|\mathbf{Y}|)] = 0. \end{aligned} \tag{32}$$

By virtue of (31) and (32) the result must follow by a consideration of p_2. Indeed,

$$\begin{aligned} &\lim_{x\to\infty} \frac{p_2}{P(|\mathbf{X}| > x)} = \lim_{x\to\infty} \int_{M^{-1}<|\mathbf{Y}|\le M} \frac{P(Z > x\,|\mathbf{Y})}{P(|\mathbf{X}| > x)} P(d\mathbf{Y}) \\ &= E\left[I_{(M^{-1},M)}(|\mathbf{Y}|) \sum_{i=1}^d \left(c_i^+ E[Y_i^\alpha I_{\{Y_i>0\}}] + c_i^- E\left[|Y_i|^\alpha\, I_{\{Y_i<0\}}\right]\right) \right]. \end{aligned} \tag{33}$$

In the last step of the proof we made use of Pratt's lemma (see Pratt (1960)) and Lemma 3.2. Now let $M \to \infty$ in (33) to conclude that the statement of the lemma is correct for $P(Z > x)$. The case $P(Z \le -x)$ is completely analogous. □

The proof of Lemma 3.6 goes as follows.

Proof. Calculation shows that $E\widetilde{Y}_t^{-\alpha/2} < \infty$ if and only if for some $x_0 > 0$,

$$\int_0^{x_0} P(\widetilde{Y}_t \le x^{2/\alpha}) x^{-2} dx < \infty. \tag{34}$$

If the X_t's are i.i.d. we have

$$\begin{aligned} P(\widetilde{Y}_t \le x^{2/\alpha}) &\le P\left(\max_{i=1,\ldots,n} X_i^2 \le x^{2/\alpha}\right) \\ &= P^n\left(|X| \le x^{1/\alpha}\right) \le \text{const}\, x^{n\gamma/\alpha}. \end{aligned} \tag{35}$$

The function $x^{n\gamma/\alpha-2}$ is integrable on $[0, x_0]$ if $n\gamma/\alpha > 1$, hence (34) holds. Now assume that $(X_1, \ldots, X_n)$ has a bounded density f_n in some neighborhood of the origin. We conclude from (35) that

$$P(\widetilde{Y}_t \le x^{2/\alpha}) \le \int_{\max_{i=1,\ldots,n} |y_i| \le x^{1/\alpha}} f_n(\mathbf{y})\, d\mathbf{y} \le \text{const}\, x^{n/\alpha},$$

for sufficiently small x, and so we may conclude that (34) holds for $n > \alpha$. This concludes the proof. □

Next we give the proof of Proposition 3.9.

Proof. (1) We observe that

$$\sum_{t=1}^n \varepsilon_t X_t^2 = \sum_{j=-\infty}^t Z_j \sum_{t=\max(1,j)}^n \psi_{t-j} X_t^2.$$

Write

$$\widetilde{\psi}_j = \frac{\sum\limits_{t=\max(1,j)}^n \psi_{t-j} X_t^2}{\sum\limits_{s=1}^n X_s^2}.$$

For fixed $k \le n$ we may then conclude from Lemma 3.4 that

$$\begin{aligned} P\left(\sum_{j=k}^n Z_j \widetilde{\psi}_j > x\right) \sim P(Z > x) \sum_{j=k}^n E[\widetilde{\psi}_j^{\,+}]^\alpha \\ + P(Z \le -x) \sum_{j=k}^n E[\widetilde{\psi}_j^{\,-}]^\alpha. \end{aligned}$$

Without loss of generality assume that $k \le 1$. Observe that

$$|\widetilde{\psi}_j| \le \sum_{t=1}^n |\psi_{t-j}|.$$

Then similar calculations as in Mikosch and Samorodnitsky (2000) show that

$$\lim_{k\to-\infty} \limsup_{x\to\infty} \frac{P\left(\left|\sum\limits_{j=-\infty}^{k-1} Z_j \widetilde{\psi}_j\right| > x\right)}{P(|Z| > x)} = 0.$$

This proves the asymptotics for the right tail of $\rho_{n,\varepsilon}$. The asymptotics for the left tail of $\rho_{n,\varepsilon}$ are analogous.

(2) The proof of the second part is analogous, making use of the representation

$$\sum_{t=1}^n \varphi_t X_t = \sum_{j=-\infty}^n \gamma_j \sum_{t=\max(1,j)}^n c_{t-j} X_t,$$

the moment condition (13) and, again, Lemma 3.4. □

References

Anderson, T.W., 1984. Introduction to Multivariate Statistical Analysis. Wiley, New York.

Andrews, B., Davis, R., Model identification for infinite variance autoregressive processes. Journal of Econometrics, in this issue (http://dx.doi.org/10.1016/j.jeconom.2012.08.009).

Basrak, B., Mikosch, T., Davis, R.A., 2002. Regular variation of GARCH processes. Stochastic Processes and Their Applications 99, 95–116.

Bingham, N.H., Goldie, C.M., Teugels, J.L., 1987. Regular Variation. Cambridge University Press, Cambridge (UK).

Blattberg, R., Sargent, T., 1971. Regression with non-Gaussian disturbances: some sampling results. Econometrica 39, 501–510.

Brainard, W., 1967. Uncertainty and the effectiveness of policy. American Economic Review, Papers and Proceedings 57, 411–425.

Breiman, L., 1965. On some limit theorems similar to the arc-sin law. Theory of Probability and Its Applications 10, 323–331.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, second ed. Springer, New York.

Butler, R.J., McDonald, J.B., Nelson, R.D., White, S.B., 1990. Robust and partially adaptive estimation of regression models. The Review of Economics and Statistics 72, 321–327.

Campbell, J.Y., Lo, A.W., McKinley, A.C., 1997. The Econometrics of Financial Markets. Princeton University Press, Princeton.

Campbell, J.Y., Shiller, R.J., 1991. Yield spreads and interest rate movements: a bird's eye view. Review of Economic Studies 58, 495–514.

Chow, G.C., 1984. Random and changing coefficient models. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, Vol. II. North-Holland, Amsterdam.

Cochrane, J.H., 2001. Asset Pricing. Princeton University Press, Princeton.

Davis, R.A., Resnick, S.I., 1996. Limit theory for bilinear processes with heavy tailed noise. Annals of Applied Probability 6, 1191–1210.

Diamond, D.W., Dybvig, P.H., 1983. Bank runs, deposit insurance, and liquidity. Journal of Political Economy 91, 401–419.

Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modelling Extremal Events for Insurance and Finance. Springer, Berlin.

Fama, E., 1976. Forward rates as predictors of future spot rates. Journal of Financial Economics 3, 361–377.

Fama, E.F., Bliss, R.B., 1987. The information in long maturity forward rates. The American Economic Review 77, 680–692.

Feige, E.L., Swamy, P.A.V.B., 1974. A random coefficient model of the demand for liquid assets. Journal of Money Credit and Banking 6, 241–252.

Feller, W., 1971. An Introduction to Probability Theory and Its Applications II. Wiley, New York.

Frenkel, J.A., 1993. On Exchange Rates. The MIT Press, Cambridge (US).

Haan, L. de, Resnick, S.I., 1977. Limit theory for multivariate sample extremes. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 40, 317–337.

Hallin, M., Swan, Y., Verdebout, T., Veredas, D., One step *R*-estimation with stable errors. Journal of Econometrics, in this issue (http://dx.doi.org/10.1016/j.jeconom.2012.08.016).

He, X., Jurečkova, J., Koenker, R., Portnoy, S., 1990. Tail behavior of regression estimators and their breakdown points. Econometrica 58, 1195–1214.

Hill, J.B., Renault, E., 2010. Generalized method of moments with tail trimming. Mimeo University of North Carolina–Chapel Hill.

Hodrick, R.J., 1987. The Empirical Evidence on the Efficiency of Forward and Futures Foreign Exchange Markets. Harwood Publishers, Chur.

Jurečkova, J., Koenker, R., Portnoy, S., 2001. Tail behavior of the least squares estimator. Statistics & Probability Letters 55, 377–384.

Lettau, M., Ludvigson, S., 2001. Resurrecting the (C)CAPM: a cross sectional test when risk premia are time-varying. Journal of Political Economy 109, 1238–1288.

Lewis, K., 1995. Puzzles in international financial markets. In: Grossman, G.M., Rogoff, K. (Eds.), Handbook of International Economics, Vol. III. North-Holland, Amsterdam, pp. 1913–1971.

Mankiw, N.G., Miron, J.A., 1986. The changing behavior of the term structure of interest rates. The Quarterly Journal of Economics 51, 211–228.

Mikosch, T., 2003. Modeling dependence and tails of financial time series. In: Finkenstädt, B., Rootzén, H. (Eds.), Extreme Value Theory and Applications. Chapman and Hall, Chichester, pp. 185–286.

Mikosch, T., Samorodnitsky, G., 2000. The supremum of a negative drift random walk with dependent heavy-tailed steps. Annals of Applied Probability 10, 1025–1064.

Nolan, J.P., Revah, D.O., Linear and nonlinear regression with stable errors. Journal of Econometrics, in this issue (http://dx.doi.org/10.1016/j.jeconom.2012.08.008).

Pratt, J., 1960. On interchanging limits and integrals. Annals of Mathematical Statistics 31, 74–77.

Resnick, S.I., 1986. Point processes, regular variation and weak convergence. Advances in Applied Probability 18, 66–138.

Resnick, S.I., 1987. Extreme Values, Regular Variation, and Point Processes. Springer, New York.

Sack, B., 2000. Does the fed act gradually? A VAR analysis. Journal of Monetary Economics 46, 229–256.

Please cite this article in press as: Mikosch, T., de Vries, C.G., Heavy tails of OLS. Journal of Econometrics (2012), doi:10.1016/j.jeconom.2012.08.015