An estimator for predictive regression: reliable inference for financial economics

Neil Shephard
Department of Economics and
Department of Statistics,
Harvard University
shephard@fas.harvard.edu

October 6, 2020

Abstract

Estimating linear regression using least squares and reporting robust standard errors is very common in financial economics, and indeed, much of the social sciences and elsewhere. For thick tailed predictors under heteroskedasticity this recipe for inference performs poorly, sometimes dramatically so. Here, we develop an alternative approach which delivers an unbiased, consistent and asymptotically normal estimator so long as the means of the outcome and predictors are finite. The new method has standard errors under heteroskedasticity which are easy to reliably estimate and tests which are close to their nominal size. The procedure works well in simulations and in an empirical exercise.

Keywords: Prediction; Regression; Robustness; Robust standard errors; Tails.

1 Introduction

Think about an outcome variable Y_1 and p predictors $\mathbf{Z}_1 = (Z_1, ..., Z_p)^\intercal$. Throughout assume $\mathrm{E}[Y_1]$ and $\mathrm{E}[\mathbf{Z}_1]$ exist (i.e. $\mathrm{E}|Y_1| < \infty$ and, element by element, $\mathrm{E}|\mathbf{Z}_1| < \infty$). Write $\mathbf{X}_1^\intercal = \left\{1, (\mathbf{Z}_1 - \mathrm{E}[\mathbf{Z}_1])^\intercal\right\}^\intercal$, where $^\intercal$ denotes a transpose, then $\mathrm{E}|\mathbf{X}_1| < \infty$. I will work with a linear in parameters "predictive regression,"

$$\mathrm{E}[Y_1|\mathbf{X}_1 = \mathbf{x}_1] = \mathbf{x}_1^\intercal \beta, \quad \text{where} \quad \mathbf{x}_1^\intercal = \left\{1, (\mathbf{z}_1 - \mathrm{E}[\mathbf{Z}_1])^\intercal\right\}^\intercal. \tag{1}$$

$\beta = (\beta_0, \beta_1, ..., \beta_p)^\intercal$ is my estimand and inference about β is my goal.

The motivation for this paper is that thick tailed predictors with heteroskedastic outcomes are very common in financial economics. Finance researchers have a strong belief in linear predictive regressions due to the theoretical backing of asset pricing linear factor models, implying a linear relation between individual stock returns and factor returns. Finance researchers nearly always assume that the mean of asset returns exists, while the vast bulk believe the variance exists. Due to the empirical evidence of their sample instability and the results from applying extreme value theory to estimate the data's tail index, many are skeptical that third or fourth moments exist (e.g. the accessible review of Cont (2001)). This challenges traditional least squares based "robust standard errors" type inference methods nearly universally used in financial economics, which rely on these higher order moments for their justification. This credibility gap rarely impacts the way applied researchers behave, perhaps understandably so because it is less than clear what action to take without potentially employing quite complicated methods. This paper provides a simple solution to this problem.

More broadly, the use of traditional least squares based robust standard errors is very common in many areas of applied statistics (e.g. see the beginning of King and Roberts (2015) for a discussion of the use in political science and MacKinnon (2012) for a discussion of the econometrics literature). The methods developed here could prove useful in other applied fields, for thick tailed data is very common, although often less apparent than in the data rich environment of financial economics.

The core of this paper focuses on the sample $(\mathbf{Z}_1, Y_1), ..., (\mathbf{Z}_n, Y_n)$, a sequence of pairs of i.i.d. random variables which each obeys (1), highlighting

$$\begin{aligned}
\widehat{\psi} &= \frac{1}{n}\sum_{j=1}^{n} \mathbf{Z}_j, \quad \mathbf{X}_j = (1, (\mathbf{Z}_j - \widehat{\psi})^\intercal)^\intercal, \\
\widehat{\beta} &= \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{j=1}^{n} S_j(\mathbf{b}), \quad S_j(\mathbf{b}) = \frac{1}{2}\left\|\mathbf{X}_j\right\|_2^{-1}\left(Y_j - \mathbf{X}_j^\intercal \mathbf{b}\right)^2, \quad \left\|\mathbf{X}_j\right\|_2 = \sqrt{\sum_{i=1}^{p+1} X_{j,i}^2}.
\end{aligned}$$

The presence of $\left\|\mathbf{X}_j\right\|_2^{-1}$ in $\widehat{\beta}$ reduces the influence of thick tailed predictors. My focus is on the role $\left\|\mathbf{X}_j\right\|_2^{-1}$ can make to allowing valid inference about β under heteroskedasticity. Downweighting extreme predictors is at the heart of the "generalized M-estimators" and "bounded-influence function" parts of the robustness literature. A classic reference to that work is Krasker and Welsch (1982).

Then

$$\frac{\partial S_j(\beta)}{\partial \beta} = -\mathbf{G}_j\left(Y_j - \mathbf{X}_j^\intercal \beta\right), \quad \mathbf{G}_j = \left\|\mathbf{X}_j\right\|_2^{-1}\mathbf{X}_j,$$

so, if the symmetric $\sum_{j=1}^{n} \mathbf{G}_j \mathbf{X}_j^{\intercal}$ is invertible, then

$$\widehat{\beta} = S_{\mathbf{G},\mathbf{X}}^{-1} S_{\mathbf{G},Y}, \quad S_{\mathbf{G},\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{G}_j \mathbf{X}_j^{\intercal}, \quad S_{\mathbf{G},Y} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{G}_j Y_j.$$

Crucially

$$\|\mathbf{G}_j\|_{\infty} = \max_{i=1,\ldots,p+1} |G_{j,i}| \leq 1,$$

which will drive the robustness of $\widehat{\beta}$ to thick tailed predictors.

$\widehat{\beta}$ will be conditionally (on the predictors) unbiased, consistent, asymptotically normal with a variance which can be estimated by

$$\frac{1}{n} S_{G,\mathbf{X}}^{-1} S_{\widehat{U}^2\mathbf{G},\mathbf{G}} S_{G,\mathbf{X}}^{-1}, \quad S_{\widehat{U}^2\mathbf{G},\mathbf{G}} = \frac{1}{n}\sum_{j=1}^{n} 1_{\||\mathbf{X}_1|/\mathrm{E}|\mathbf{X}_1|\|_{\infty} \leq cn^{1/5}} \widehat{U}_j^2 \mathbf{G}_j \mathbf{G}_j^{\intercal} \tag{2}$$

where $U_j = Y_j - \mathbf{X}_j^{\intercal}\widehat{\beta}$, so long as $\mathrm{Var}(U_1) < \infty$ and $\mathrm{E}|\mathbf{X}_1| < \infty$. In practice I take $c = 10$, so the truncation $1_{\||\mathbf{X}_1|/\mathrm{E}|\mathbf{X}_1|\|_{\infty} \leq cn^{1/5}}$ is irrelevant except for the most extraordinary predictors. Without the truncation we need the additional condition that $\mathrm{Var}(\mathbf{X}_1) < \infty$.

In comparison, Assumption 4 of White (1980) spells out that Eicker (1967), Huber (1967) and White (1980) robust standard errors needs $\mathrm{Var}(\mathbf{X}_1^2) < \infty$ for inference on β based on least squares

$$\widehat{\beta}_{LS} = S_{\mathbf{X},\mathbf{X}}^{-1} S_{\mathbf{X},Y}, \quad S_{\mathbf{X},\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_j \mathbf{X}_j^{\intercal}, \quad S_{\mathbf{X},Y} = \sum_{j=1}^{n} \mathbf{X}_j Y_j,$$

to be asymptotically valid. Recall these standard errors are based on

$$\frac{1}{n} S_{\mathbf{X},\mathbf{X}}^{-1} \mathbf{S}_{\widehat{U}_{LS}^2\mathbf{X},\mathbf{X}} S_{\mathbf{X},\mathbf{X}}^{-1}, \quad \mathbf{S}_{\widehat{U}_{LS}^2\mathbf{X},\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n} \widehat{U}_{LS,j}^2 \mathbf{X}_j \mathbf{X}_j^{\intercal}, \quad \widehat{U}_{LS,j} = (Y_j - \mathbf{X}_j^{\intercal}\widehat{\beta}_{LS}).$$

Unfortunately, financial economists may well not have those four moments available to them. Monte Carlo and empirical results presented later will demonstrate this asymptotic worry is important in practice. Further, the results suggest that even if more moments exist than four the finite sample inference is still very fragile unless n is very substantial. Overall, in my opinion, the evidence suggests Eicker (1967), Huber (1967) and White (1980) type "robust standard errors" are not credible in financial economics. $\widehat{\beta}$ is one potential solution.

Before ending this introduction, it is helpful to take a step back to be clear about some labels. Under assumption (1) β is equal to $\gamma = \mathrm{Var}(\mathbf{X}_1)^{-1}\mathrm{Cov}(\mathbf{X}_1, Y_1)$, so long as the relevant moments exist. The "descriptive regression" γ is just a rewrite of centered second moments of $(\mathbf{X}_1, Y_1)$ and is well defined (if the moments exist) even if (1) does not hold. In financial economics, "predictive regression" usually implies Y_1 is in the future compared to $\mathbf{X}_1$, e.g. Stambaugh (1999). I am using a broader statistical/machine learning terminology: prediction is where I have seen $\mathbf{X}_1$ and wish to learn Y_1 (if time is involved it would a "forecasting regression"). "Regression" has many meanings and it is crucial to understand which is being used and why.

The remaining parts of this paper are as follows. In Section 2 I will focus on a scalar predictor and explain where $\widehat{\beta}$ comes from and derive its main inferential properties. While

doing this I will review the literature on this topic, linking results across different intellectual fields.

In Section 3 I provide conditions for identifying β and derive a corresponding method of moments estimator $\widehat{\beta}$. Section 4.1 holds the main condition properties of $\widehat{\beta}$, conditioning on the predictors. Section 4.2 contains the corresponding unconditional properties of $\widehat{\beta}$. In both sections $\widehat{\beta}$ is compared to the corresponding least squares estimator $\widehat{\beta}_{LS}$. Section 5 presents the results from various simulation experiments to see how effective the asymptotics guidance is.

Section 6 contains results from a massive number of hypothesis tests using $\widehat{\beta}$, where I identify stocks with high betas or low betas. This allows me to form high (or low) beta portfolios, which is a potentially useful investment vehicle for investors unable to take on financial leverage (e.g. young savers into pensions). I also study how these procedures work as they are rolled through the time series database.

I state my conclusions in Section 7. Any lengthy proof of a Theorem stated in the main text is given in the Appendix.

2 Why is $\widehat{\beta}$ interesting and the literature

The main virtues of $\widehat{\beta}$ are seen in the most stripped down case: the focus of this section.

Assume a linear in parameters "predictive regression"

$$\mathrm{E}[Y_1|X_1 = x_1] = \beta_1 x_1, \tag{3}$$

for outcome Y_1 and scalar predictor $X_1 = Z_1$, where $\mathrm{E}[Y_1] = \mathrm{E}[Z_1] = 0$. As each item is a scalar, no bolding will be used here. Upper cases denote random variables, lower cases fixed numbers.

Then,

$$\|x_1\|_2 = |x_1|, \qquad g_1 = \|x_1\|_2^{-1} x_1 = sign(x_1),$$

so

$$\widehat{\beta}_1 = \underset{b_1}{\operatorname{argmin}} \sum_{j=1}^{n} S_j(b_1), \quad S_j(b_1) = \frac{1}{2}|x_j|^{-1}\left(y_j - x_j b_1\right)^2, \quad \frac{\partial S_j(b_1)}{\partial b_1} = -sign(x_j)\left(y_j - x_j b_1\right),$$

implying

$$\widehat{\beta}_1 = S_{G,X}^{-1} S_{G,Y} = \frac{\sum_{j=1}^{n} sign(x_j) y_j}{\sum_{j=1}^{n} |x_j|},$$

where $S_{G,X} = \frac{1}{n}\sum_{j=1}^{n} g_j x_j = \frac{1}{n}\sum_{j=1}^{n} |x_j|$ and $S_{G,Y} = \frac{1}{n}\sum_{j=1}^{n} g_j y_j = \frac{1}{n}\sum_{j=1}^{n} sign(x_j) y_j$.

I give nine features of $\widehat{\beta}_1$, weaving them together with a literature review.

2.1 Ways of deriving $\widehat{\beta}_1$

The first three features are different ways of deriving $\widehat{\beta}_1$.

First, multiply both sides of (3) by g_1, then

$$g_1 \mathrm{E}[Y_1|X_1 = x_1] = \beta_1 sign(x_1)x_1 = \beta_1|x_1|.$$

If $\mathrm{E}[G_1X_1]$ and $\mathrm{E}[G_1Y_1|$ exist, then unconditionally

$$\mathrm{E}[G_1Y_1] = \beta_1 \mathrm{E}[G_1X_1].$$

Crucially $|G_1| \leq 1$, so a sufficient condition for $\mathrm{E}[G_1Y_1|$ to exist is that $\mathrm{E}|Y_1| < \infty$. The same argument implies $\mathrm{E}[G_1Z_1]$ exists if $\mathrm{E}|X_1| < \infty$. If, in addition, $\mathrm{E}[G_1Z_1] = \mathrm{E}|X_1| > 0$ then

$$\beta_1 = \frac{\mathrm{E}[G_1Y_1]}{\mathrm{E}[G_1Z_1]} = \frac{\mathrm{E}[sign(Z_1)Y_1]}{\mathrm{E}|Z_1|}.$$

So a sufficient condition for β_1 to be identified is $0 < \mathrm{E}|Z_1| < \infty$ and $\mathrm{E}|Y_1| < \infty$. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be a sequence of pairs of random variables which each obeys (3). Then

$$\widehat{\beta}_1 = \frac{\sum_{j=1}^n sign(X_j)Y_j}{\sum_{j=1}^n |X_j|}, \tag{4}$$

is a method of moments estimator. By the strong law of large numbers, under just two conditions, $0 < \mathrm{E}|X_1| < \infty$ and $\mathrm{E}|Y_1| < \infty$,

$$\widehat{\beta}_1 \xrightarrow{p} \frac{\mathrm{E}[sign(X_1)Y_1]}{\mathrm{E}|X_1|} = \beta_1.$$

Hence $\widehat{\beta}_1$ is consistent if the data is a tad less thick tailed than, for example, Cauchy random variables.

Second, $\widehat{\beta}_1$ is an Instrumental Variable (IV) estimator, where the "instruments" are $sign(X_j)$. Often, IV estimators behave notoriously poorly in many of their applications as the "relevance" condition of instrumental variables is "weak" (e.g. the reviews in Andrews et al. (2019)). This is not the case here, as the relevance condition $\mathrm{E}[sign(X_1)X_1] = \mathrm{E}|X_1| > 0$ should hold strongly.

Third, $\widehat{\beta}_1$ is the maximum quasi-likelihood (ML) estimator from the contrived model $Y_j|X_j = x_j \overset{indep}{\sim} N(\beta_1 x_j, |x_j|\sigma^2)$. This implies the existence of a quasi-likelihood $\log L(\beta_1) = \frac{\beta_1}{\sigma^2}\sum_{j=1} y_j sign(x_j) - \frac{1}{2\sigma^2}\beta_1^2 \sum_{j=1} |x_j|$, which downweights predictors with very large $|x_j|$ compared to the tradition homoskedastic quasi-likelihood case. This structure also implies that $\widehat{\beta}_1$ is a weighted least squares estimator, where the weights are $\{|x_j|\sigma^2\}^{-1}$, and so is a special case of generalized least squares.

My fourth point is different. Divide the top and bottom of (4) by n and write

$$\widehat{\beta}_1 = \frac{\overline{1_{X>0}Y} - \overline{1_{X<0}Y}}{\overline{1_{X>0}X} - \overline{1_{X<0}X}}, \quad \text{where, e.g.,} \quad \overline{1_{X>0}Y} = \frac{1}{n}\sum_{j=1}^n 1_{X_j>0}Y_j,$$

then the geometry of the estimator is shown in Figure 1, where the slope of the green line is $\widehat{\beta}_1$. The length of horizontal red line is $\overline{1_{X>0}X} - \overline{1_{X<0}X} > 0$, while the vertical red line moves down from $\overline{1_{X>0}Y}$ to $\overline{1_{X<0}Y}$.

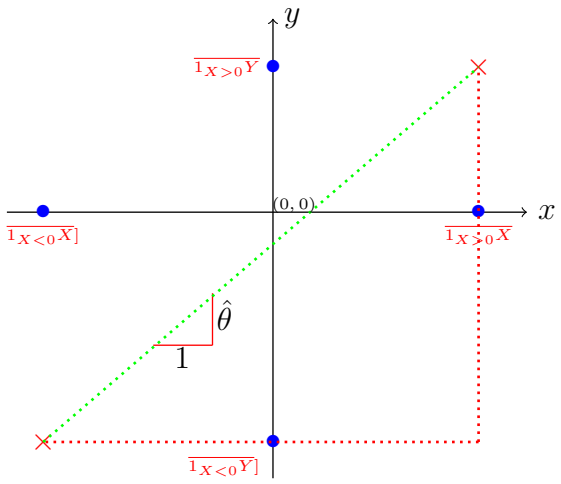

Figure 1: Slope of the green line is $\hat{\theta}$. Length of horizontal red line is $\overline{1_{X>0}X} - \overline{1_{X<0}X} > 0$, the vertical red line moves down from $\overline{1_{X>0}Y}$ to $\overline{1_{X<0}Y}$.

2.2 Major properties of $\widehat{\beta}_1$

The next two features are the main inferential properties of $\widehat{\beta}_1$.

Fifth, in terms of conditional inference, if the pairs (X_j, Y_j) are independent and (3) holds for each j, then $\mathrm{E}[\widehat{\beta}_1|(\mathbf{X} = \mathbf{x})] = \beta_1$, where $\mathbf{X} = (X_1, ..., X_n)$ and the observed predictors $x = (x_1, ..., x_n)$. Further, for $j = 1, ..., n$, if $\sigma_j^2(x_j) = \mathrm{Var}(Y_j|X_j = x_j) < \infty$,

$$\mathrm{Var}[\widehat{\beta}_1|\,(\mathbf{X} = \mathbf{x})] = \frac{\sum_{j=1}^n \sigma_j^2(x_j)}{\left(\sum_{j=1}^n |x_j|\right)^2} = \frac{1}{n}\frac{\frac{1}{n}\sum_{j=1}^n \sigma_j^2(x_j)}{\left(\frac{1}{n}\sum_{j=1}^n |x_j|\right)^2}.$$

Then $\frac{1}{n}\sum_{j=1}^n \sigma_j^2(x_j)$ can be estimated by $\frac{1}{n}\sum_{j=1}^n (Y_j - \widehat{\beta}_1 x_j)^2$ (this will be discussed in more detail shortly). This makes inference based on the approximate pivot

$$\widehat{T}_{\widehat{\beta}} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sum_{j=1}^n (Y_j - \widehat{\beta}_1 x_j)^2}{\left(\sum_{j=1}^n |x_j|\right)^2}}}$$

effective for thick tailed heteroskedastic data. Whether the researcher assumes homoskedasticity or not does not change the form of $\widehat{T}_{\widehat{\beta}}$. It was this property which initially made me interested in $\widehat{\beta}_1$.

Sixth, in terms of unconditional inference, if the pairs (X_j, Y_j) are independent and identically distributed (i.i.d.), then the strong law of large numbers implies that, as $n \to \infty$,

$$\widehat{\beta}_1 \xrightarrow{p} \frac{\mathrm{E}[sign(X_1)Y_1]}{\mathrm{E}|X_1|} = \frac{\mathrm{E}[1_{X_1>0}Y_1] - \mathrm{E}[1_{X_1<0}Y_1]}{\mathrm{E}[1_{X_1>0}X_1] - \mathrm{E}[1_{X_1<0}X_1]} = \beta^*,$$

so long as $\mathrm{E}|Y_1| < \infty$ and $0 < \mathrm{E}|X_1| < \infty$. Here, β_1^* is a "pseudo-true" value of β. If (3) holds, then Adam's Law implies that $\mathrm{E}[1_{X_1>0}Y_1] = \beta_1 \mathrm{E}[1_{X_1>0}X_1]$ and $\mathrm{E}[1_{X_1<0}Y_1] = \beta_1 \mathrm{E}[1_{X_1<0}X_1]$,

so $\beta_1^* = \beta_1$, forcing $\widehat{\beta}_1 \xrightarrow{p} \beta_1$. Further, defining $U_1 = Y_1 - X_1\beta_1$ and additionally assuming $\text{Var}(U_1) < \infty$, then unconditionally

$$\sqrt{n}(\widehat{\beta}_1 - \beta_1) \xrightarrow{d} N\left(0, \frac{\text{Var}(U_1)}{\{\text{E}|X_1|\}^2}\right),$$

hence heteroskedasticity has no impact on the limit distribution of $\widehat{\beta}_1$. Further, $\text{E}|X_1|$ can be estimated by $\frac{1}{n}\sum_{j=1}^n |X_j|$. Finally, define $\widehat{U}_j = Y_j - X_j\widehat{\beta}_1 = U_j - X_j(\widehat{\beta}_1 - \beta_1)$ where $U_j = Y_j - X_j\beta_1$, so

$$\widehat{\text{Var}(U_1)} = \frac{1}{n}\sum_{j=1}^n \widehat{U}_j^2 = \frac{1}{n}\sum_{j=1}^n U_j^2 + (\widehat{\beta}_1 - \beta_1)^2\frac{1}{n}\sum_{j=1}^n X_j^2 - 2(\widehat{\beta}_1 - \beta_1)\frac{1}{n}\sum_{j=1}^n U_j X_j.$$

If $\text{Var}(X_1) < \infty$ exists then $\text{Var}(U_1)$ can be consistently estimated by $\widehat{\text{Var}(U_1)}$. More broadly, if $\text{Var}(X_1)$ does not exist then $\widehat{\text{Var}(U_1)}$ performs poorly in theory and in simulations (this will be reported in Sections 4.3 and 5). Instead, a weighted version, which clips the estimator for large absolute predictors, $\frac{1}{n}\sum_{j=1}^n \widehat{U}_j^2 w(X_j)$ where the weight $w(x) = 1_{\frac{|x|}{\text{E}|X_1|} < dn^{1/5}}$, is consistent for $\text{Var}(U_1)$, requiring, again just requiring $\text{E}|X_1| < \infty$. In simulations, I take $d = 10$ (so if $n = 10$ then $dn^{1/5} \simeq 16$), so the clipping will have literally no impact on nearly all applied work. However, the evidence suggests the weight is a worthwhile guardrail for very thick tailed data.

2.3 Relating $\widehat{\beta}$ to other estimators

Seventh, in the context of linear regressions for stable random variables, Blattberg and Sargent (1971) derived

$$\widetilde{\beta}_1 = \sum_{j=1}^n |x_j|^c sign(x_j) Y_j / \sum_{j=1}^n |x_j|^{1+c}, \quad c > 0,$$

regarding the predictors as non-stochastic, minimizing the α-stable scale in the class of linear unbiased estimators. When $c = 0$, $\widetilde{\beta}_1$ would be $\widehat{\beta}_1$, but they did not cover that case, nor its analytic properties. When $c > 0$ standard errors are not robust to heteroskedasticity. Samorodnitsky et al. (2007) studied the distributional properties of $\widetilde{\beta}_1$ under very heavy tails in X_1, but assuming independence between the predictors and the regression errors. This independence assumption takes them outside our interests. Gorji and Aminghafari (2019) builds on Blattberg and Sargent (1971) and Samorodnitsky et al. (2007) towards non-parametric regression.

Eighth, So and Shin (1999) studied estimating autoregressions with the instrument $sign(Y_{j-1})$, yielding an estimator

$$\overline{\beta}_1 = \sum_{j=2}^n sign(Y_{j-1}) Y_j / \sum_{j=2}^n |Y_{j-1}|.$$

To So and Shin (1999), $\overline{\beta}_1$ had attractive properties in heavy tailed time series. They call this a "Cauchy estimator", after Cauchy (1836) (who followed up his first paper with 6 others

around this topic), noting that $\overline{\beta}_1$ can be thought of as an instrumental variable estimator and a generalized least squares estimator. The historians of least squares and regression in statistics usually associate Cauchy's work with numerical interpolation (e.g. Seal (1967), Ch. 13 of Farebrother (1999) and Ch. 4 of Heyde and Seneta (1977)). It matches $\widehat{\beta}_1$ only in the scalar case with no intercept. Heyde and Seneta (1977) detail Cauchy's work on regression from a modern perspective. Ch. 13.5 of Linnik (1961) discusses the multivariate Cauchy's method, proving it is unbiased and derives the variance of $\widehat{\beta}_1$ in the scalar case for non-stochastic predictors under homoskedasticity.

Phillips et al. (2004) generalized the So and Shin (1999) use of $sign(Y_{j-1})$ in an autoregression to an "instrument generating function" $F(Y_{j-1})$, where F is an asymptotically homogenous function. Kim and Meddahi (2020) mention the So and Shin (1999) approach to fitting autoregressions in the context of time series of realized volatility type objects (which tend to be thick tailed). They also link to Samorodnitsky et al. (2007). Their interest was in consistent estimation for time series with heavy tailed regression errors. Ibragimov et al. (2020) look at time series estimators of $\overline{\beta}_1$ type under volatility clustering.

Mikosch and de Vries (2013) studied the properties of least squares under heavy tailed predictors, providing interesting results and references. Hill and Renault (2010) devised trimming methods for the GMM which allow for both heavy tailed variables and Gaussian limit theory. Hallin et al. (2010) looked at using rank based methods in regression context with heavy tailed data, while Butler et al. (1990) use adaptive statistical models for regressions, assuming predictors and prediction errors are independent.

Ninth, more broadly, Balkema and Embrechts (2018) provides a review of a substantial literature on robust estimation and heavy tailed data, as well as comparing procedures using Monte Carlo methods. Related work is Kurz-Kim and Loretan (2014). Nolan and Ojeda-Revah (2013) looks at linear regressions with heavy tailed errors. Of course, these last two papers interface with the influential robustness literature, reviewed by Hampel et al. (2005). Sun et al. (2020) is an interesting recent paper which is close to our setup. It uses a Huber loss for a linear predictive regression where the threshold is selected adaptively so that asymptotically they still recover the estimand, the parameter indexing the predictive regression. The Sun et al. (2020) procedure is likely to be asymptotically more efficient than $\widehat{\beta}_1$.

The use of

$$\widehat{\beta} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{j=1}^{n} \|\mathbf{X}_j\|_2^{-1} \left(Y_j - \mathbf{X}_j^{\intercal}\mathbf{b}\right)^2,$$

is a special case of the important wide ranging work on the bounded influence function literature, which goes back to Mallows (1975a,b) and Krasker and Welsch (1982), often under the label "generalized M-estimators". Most of their focus is on efficient robust estimation of β. My interest is in being able to estimate the standard errors under heteroskedasticity so that inference in financial economics is reliable. I am also keen not on changing the estimand, which rules out much of modern robust regression.

2.4 Comparing $\widehat{\beta}$ to least squares

I now compare $\widehat{\beta}_1$ to the ML estimator from the conditionally Gaussian linear regression $Y_j|X_j = x_j \overset{indep}{\sim} N(\beta_1 x_j, \sigma^2)$,

$$\widehat{\beta}_{LS,1} = \frac{\sum_{j=1}^n X_j Y_j}{\sum_{j=1}^n X_j^2},$$

the celebrated "least squares" estimator. Of course, under homoskedasticity, $\widehat{\beta}_{LS,1}$ will be more efficient than $\widehat{\beta}_1$. For i.i.d. pairs, famously, if enough moments exist, then unconditionally

$$\sqrt{n}(\widehat{\beta}_{LS,1} - \beta_1) \xrightarrow{d} N\left(0, \frac{\mathrm{Var}(X_1 U_1)}{\left\{\mathrm{E}[X_1^2]\right\}^2}\right).$$

Then $\frac{1}{n}\sum_{j=1}^n X_j^2$ estimates $\mathrm{E}[X_1^2]$. Define $\widehat{U}_{LS,j} = Y_j - X_j\widehat{\beta}_{LS,1} = U_j - X_j(\widehat{\beta}_{LS,1} - \beta_1)$ where, again, $U_j = Y_j - X_j\beta_1$. Then

$$\widehat{\mathrm{Var}(X_1 U_1)} = \frac{1}{n}\sum_{j=1}^n X_j^2 \widehat{U}_{LS,j}^2 = \frac{1}{n}\sum_{j=1}^n X_j^2 U_j^2 + (\widehat{\beta}_{LS,1} - \beta_1)^2 \frac{1}{n}\sum_{j=1}^n X_j^4 - 2(\widehat{\beta}_1 - \beta_1)\frac{1}{n}\sum_{j=1}^n U_j X_j^3,$$

which needs $\mathrm{E}(X_1^4) < \infty$ to behave well as an estimator of $\mathrm{Var}(X_1 U_1)$, in theory and in simulations. If enough moments exist, then, taken together, this motivates robust standard errors based on $\frac{1}{n}\sum_{j=1}^n X_j^2 \widehat{U}_{LS,j}^2$, following Eicker (1967), Huber (1967) and White (1980) (where Assumption 4 spells out the need for $\mathrm{E}(X_1^4) < \infty$). Robust standard errors are used in vast numbers of applied papers. Unfortunately $\widehat{\mathrm{Var}(X_1 U_1)}$ is a poor estimator unless (i) n is very large, (ii) U_1 is known to be independent of X_1 or (iii) the predictors are thin tailed. This makes valid inference based on the asymptotic pivot

$$\widehat{T}_{LS} = \frac{\widehat{\beta}_{LS,1} - \beta_1}{\sqrt{\frac{\sum_{j=1}^n X_j^2 (Y_j - \widehat{\beta} X_j)^2}{\left(\sum_{j=1}^n X_j^2\right)^2}}},$$

challenging for data in finance. It is well known that $\widehat{T}_{LS}$ often has poor finite sample properties, although the infeasible version of this $T_{LS} = (\widehat{\beta}_{LS,1} - \beta_1)/\sqrt{\mathrm{Var}(\widehat{\beta}_{LS,1}|(\mathbf{X} = \mathbf{x}))}$ does not. Some try to mend this problem using a bootstrap of the approximate pivot $\widehat{T}_{LS}$ or an Edgeworth expansion, e.g. MacKinnon and White (1985), Hall (1992), MacKinnon (2012) and Hausman and Palmer (2012).

As I said, under homoskedasticity $\widehat{\beta}_{LS,1}$ will be more efficient than $\widehat{\beta}_1$ (e.g. the Gauss-Markov Theorem or, in the Gaussian outcomes case, Cramér-Rao inequality), with

$$\frac{\mathrm{Var}(\widehat{\beta}_1)}{\mathrm{Var}(\widehat{\beta}_{LS,1})} \simeq \frac{\{\mathrm{E}[X_1^2]\}}{\{\mathrm{E}|X_1|\}^2} \geq 1,$$

by Jensen's inequality. This point goes back at least to Ch. 13.5 of Linnik (1961). If $X_1 \in \{-1, 1\}$, with equal probability, then $\mathrm{E}[X_1^2]/\{\mathrm{E}|X_1|\}^2 = 1$, so $\widehat{\beta}$ is fully efficient. Of course it

is, $sign(X_j) = X_j$ in that case, so $\widehat{\beta}_1 = \widehat{\beta}_{LS,1}$. Under $X_1 \sim N(0, \lambda^2)$, then $\mathrm{E}[X_1^2]/\left\{\mathrm{E}|X_1|\right\}^2 = \pi/2 \simeq 1.57$ hence the $\widehat{\beta}_{LS,1}$ is substantially more efficient than $\widehat{\beta}_1$. Under the thicker tailed $X_1 \sim Laplace(0, \lambda)$, so $\mathrm{E}[X_1^2]/\left\{\mathrm{E}|X_1|\right\}^2 = 2$. If X_1 is very thick tailed, then $\mathrm{E}[X_1^2]$ can go to infinity in cases where $\mathrm{E}|X_1|$ is finite. Then $\widehat{\beta}_{LS,1}$ is a much more precise estimator, on average, but the Gaussian CLT no longer holds for $\widehat{\beta}_{LS,1}$ in cases where the CLT for $\widehat{\beta}_1$ is still useful. This suggests CLT for $\widehat{\beta}_1$ may be a more practical guide in the kind of thick tailed data often seen in finance, for example.

3 Identification and estimation of β

Again, think about an outcome variable Y_1 and p predictors $\mathbf{Z}_1 = (Z_1, ..., Z_p)^{\mathsf{T}}$, where $\mathrm{E}\,|Y_1| < \infty$ and $\mathrm{E}\,|\mathbf{Z}_1| < \infty$. The following is enough to establish identification of β.

Assumption 1 (Joint law of $(\mathbf{Z}_1, Y_1)$) *A1.* $\mathrm{E}\,|Y_1| < \infty$ *and* $\mathrm{E}\,|\mathbf{Z}_1| < \infty$. *Write* $\psi = \mathrm{E}[\mathbf{Z}_1]$,

$$\mathbf{X}_1(\psi)^{\mathsf{T}} = \left\{1, (\mathbf{Z}_1 - \psi)^{\mathsf{T}}\right\}^{\mathsf{T}}, \quad \textit{and} \quad \mathbf{G}_1(\psi) = \|\mathbf{X}_1(\psi)\|_2^{-1}\,\mathbf{X}_1(\psi).$$

A2. There exists a single β *such that*

$$\mathrm{E}[Y_1|\mathbf{Z}_1 = \mathbf{z}_1] = \mathbf{x}_1^{\mathsf{T}}\beta, \quad \mathbf{x}_1^{\mathsf{T}} = \left\{1, (\mathbf{z}_1 - \psi)^{\mathsf{T}}\right\}^{\mathsf{T}} \tag{5}$$

so all $\mathbf{z}_1$.

A3.

$$\mathrm{E}\left[\mathbf{G}_1(\psi)\mathbf{X}_1(\psi)^{\mathsf{T}}\right] = \mathrm{E}\left[\|\mathbf{X}_1(\psi)\|_2^{-1}\,\mathbf{X}_1(\psi)\mathbf{X}_1(\psi)^{\mathsf{T}}\right]$$

is positive definite.

As A1 includes an intercept, note that $\|\mathbf{X}_1(\psi)\|_2 \geq 1$, while $\|\mathbf{G}_1(\psi)\|_\infty \leq 1$.

Theorem 1 *Under A1,*

$$\mathrm{E}\left[\mathbf{G}_1(\psi)\mathbf{X}_1(\psi)^{\mathsf{T}}\right]$$

exists and is symmetric, positive semidefinite. Under A1+A3, it is also positive definite.

Proof. $\|\mathbf{G}_1(\psi)\|_\infty \leq 1$ so A1 implies $\mathrm{E}\left[\mathbf{G}_1(\psi)\mathbf{X}_1(\psi)^{\mathsf{T}}\right]$ exists. By construction, it is symmetric, positive semi-definite. Assumption A3 pushes this to positive definite. QED.

The estimand will be β, while $\psi = \mathrm{E}[\mathbf{Z}_1]$ will be a "nuisance". Using Theorem 1, the following is straightforward.

Theorem 2 (Identification) *Assume A1-A3, then*

$$\begin{aligned} \psi &= \mathrm{E}[\mathbf{Z}_1] \\ \beta &= \left\{\mathrm{E}\left[\mathbf{G}_1(\psi)\mathbf{X}_1(\psi)^{\mathsf{T}}\right]\right\}^{-1}\mathrm{E}[\mathbf{G}_1(\psi)Y_1]. \end{aligned} \tag{6}$$

This Theorem says that β and ψ can be uniquely determined, that is, identified, from $\mathrm{E}[\mathbf{Z}_1]$, $\mathrm{E}[\mathbf{G}_1(\psi)Y_1]$ and $\mathrm{E}\left[\mathbf{G}_1(\psi)\mathbf{X}_1(\psi)^{\intercal}\right]$. Further, and crucially, all three of these terms are guaranteed to exist if both $\mathrm{E}[\mathbf{Z}_1]$ and $\mathrm{E}[Y_1]$ exist. Adding Assumption A3 is the only substantial assumption made beyond the core model A1 and A2.

Now turn to estimation.

Let $(\mathbf{Z}_1, Y_1), ..., (\mathbf{Z}_n, Y_n)$ be a sequence of pairs of random variables which each obeys A1-A3.

Define

$$\begin{aligned}\overline{\mathbf{Z}} &= \frac{1}{n}\sum_{j=1}^{n}\mathbf{Z}_j, \quad \mathbf{X}_j = (1, (\mathbf{Z}_j - \overline{\mathbf{Z}})^{\intercal})^{\intercal}, \quad \mathbf{G}_j = \|\mathbf{X}_j\|_2^{-1}\mathbf{X}_j, \\ S_{\mathbf{G},\mathbf{X}} &= \frac{1}{n}\sum_{j=1}^{n}\mathbf{G}_j\mathbf{X}_j^{\intercal}, \quad S_{\mathbf{G},Y} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{G}_j Y_j.\end{aligned}$$

Importantly, $\|\mathbf{G}_j\|_\infty \leq 1$ and $S_{\mathbf{G},\mathbf{X}}$ is symmetric and positive semi-definite.

We now introduce a method of moment estimator of (ψ,β).

Definition 1 *Assume $S_{\mathbf{G},\mathbf{X}}$ is positive definite. Using the moment condition (6), define a method of moment estimator*

$$\widehat{\psi} = \overline{\mathbf{Z}}, \quad \textit{and} \quad \widehat{\beta} = S_{\mathbf{G},\mathbf{X}}^{-1}S_{\mathbf{G},Y}.$$

$\widehat{\beta}$ is an instrumental variable (IV) estimator, that uses the $\mathbf{G}_j$ as instruments (although note that $\widehat{\psi}$ is buried within $\mathbf{G}_j$).

Example 1 *If $p = 1$ and no intercept, so $X_j = Z_j - \overline{Z}$, then $\mathbf{G}_j = \|\mathbf{X}_j\|_2^{-1}\mathbf{X}_j = sign(Z_j - \overline{Z})$,*

$$\widehat{\beta} = \frac{\sum_{j=1}^{n} sign(Z_j - \overline{Z})Y_j}{\sum_{j=1}^{n}|Z_j - \overline{Z}|},$$

which is centered version of the estimator discussed in Section 2.

It is sometimes helpful to unpack β into its elements $\beta = (\beta_0, \beta_{1:p}^{\intercal})^{\intercal}$.

Theorem 3 *Assume $S_{\mathbf{G},\mathbf{X}} > 0$. Define the weights and weighted averages*

$$w_j = \frac{\|\mathbf{X}_j\|_2^{-1}}{\sum_{i=1}^{n}\|\mathbf{X}_i\|_2^{-1}}, \quad \widetilde{Y} = \sum_{j=1}^{n} w_j Y_j, \quad \widetilde{\mathbf{Z}} = \sum_{j=1}^{n} w_j \mathbf{Z}_j,$$

and the weighted sums of outer products

$$\widetilde{S}_{\mathbf{Z}-\widetilde{\mathbf{Z}},\mathbf{Z}-\widetilde{\mathbf{Z}}} = \sum_{j=1}^{n} w_j \left(\mathbf{Z}_j - \widetilde{\mathbf{Z}}\right)\left(\mathbf{Z}_j - \widetilde{\mathbf{Z}}\right)^{\intercal}, \quad \widetilde{S}_{\mathbf{Z}-\widetilde{\mathbf{Z}},Y-\widetilde{Y}} = \sum_{j=1}^{n} w_j \left(\mathbf{Z}_j - \widetilde{\mathbf{Z}}\right)\left(Y_j - \widetilde{Y}\right).$$

Then

$$\widehat{\beta}_{1:p} = \widetilde{S}_{\mathbf{Z}-\widetilde{\mathbf{Z}},\mathbf{Z}-\widetilde{\mathbf{Z}}}^{-1}\widetilde{S}_{\mathbf{Z}-\widetilde{\mathbf{Z}},Y-\widetilde{Y}}, \quad \widehat{\beta}_0 = \widetilde{Y} - \left(\widetilde{\mathbf{Z}} - \overline{\mathbf{Z}}\right)^{\intercal}\widehat{\gamma},$$

delivering the j-th residual $\widehat{U}_j = \left(Y_j - \widehat{\beta}_0\right) - (Z_j - \overline{\mathbf{Z}})^{\intercal}\widehat{\beta}_{1:p}$, $j = 1, ..., n$.

Proof. Given in the Appendix.

Example 2 *If $p = 1$ then*

$$\widetilde{Y} = \frac{\sum_{j=1}^{n} \frac{Y_j}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}}{\sum_{j=1}^{n} \frac{1}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}}, \quad \widetilde{Z} = \frac{\sum_{j=1}^{n} \frac{Z_j}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}}{\sum_{j=1}^{n} \frac{1}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}}, \quad \widehat{\beta}_0 = \widetilde{Y} - \left(\widetilde{Z} - \overline{Z}\right)\widehat{\beta}_1$$

$\widehat{\beta}_{0,LS} = \overline{Y}$ *and*

$$\widehat{\beta}_1 = \frac{\sum_{j=1}^{n} \frac{\left(Z_j-\widetilde{Z}\right)\left(Y_j-\widetilde{Y}\right)}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}}{\sum_{j=1}^{n} \frac{\left(Z_j-\widetilde{Z}\right)^2}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}}, \quad \widehat{\beta}_{1,LS} = \frac{\sum_{j=1}^{n}\left(Z_j - \overline{Z}\right)Y_j}{\sum_{j=1}^{n}\left(Z_j - \overline{Z}\right)^2}.$$

4 Properties of $\widehat{\beta}$

4.1 Conditional properties of $\widehat{\beta}$

There are two broad ways of performing inference on the parameters that index a predictive regression: conditionally and unconditionally. First, focus on the conditional case.

Assumption 2 (Conditional assumptions) *B1. The matrix*

$$S_{\mathbf{G},\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{G}_j \mathbf{X}_j^{\intercal}$$

is positive definite.

B2. The pairs $(\mathbf{Z}_1, Y_1), ..., (\mathbf{Z}_n, Y_n)$ are independent.

B3. $\mathrm{Var}(Y_j|\mathbf{Z}_j = \mathbf{z}_j) = \sigma_j^2 < \infty$*, for* $j = 1, 2..., n$*.*

B4. The matrix

$$S_{\sigma^2\mathbf{G},\mathbf{G}} = \frac{1}{n}\sum_{j=1}^{n} \sigma_j^2 \mathbf{G}_j \mathbf{G}_j^{\intercal}$$

is positive definite.

B5. $\mathrm{E}[|U_j|^3|(\mathbf{Z} = \mathbf{z})] < \infty$*, for* $j = 1, 2..., n$*, where* $U_j = Y_j - \mathbf{X}_j^{\intercal}\beta$*.*

Write all the random predictors as $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)$ and their observed version in the sample as $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_n)$**.**

Theorem 4 *If B1-B2 and A1-A2 hold for each $j = 1, ..., n$, then* $\mathrm{E}[\widehat{\beta}|(\mathbf{Z} = \mathbf{z})] = \beta$**.** *If B1-B3 and A1-A2 hold, then*

$$\mathrm{Var}(\widehat{\beta}|(\mathbf{Z} = \mathbf{z})) = \frac{1}{n}\Psi_n, \quad \Psi_n = S^{-1}_{\mathbf{G},\mathbf{X}} S_{\sigma^2\mathbf{G},\mathbf{G}} S^{-1}_{\mathbf{G},\mathbf{X}}.$$

Further, under B1-B5, there exists a constant $c > 0$, such that

$$\begin{aligned} &\sup_{A \in \mathcal{C}_{p+1}} \left| \Pr(\sqrt{n}\left(\widehat{\beta} - \beta\right) \in A|(\mathbf{Z} = \mathbf{z})) - \Pr(N(0, \Psi_n) \in A|(\mathbf{Z} = \mathbf{z})) \right| \\ \leq \;& c\frac{(p+1)^{1/4}}{n^{1/2}} \left(n^{-1} \sum_{j=1}^{n} \varsigma_j \right), \end{aligned}$$

where $\mathcal{C}_{p+1}$ denotes the set of all convex subsets of R^{p+1} and

$$\varsigma_j = \mathrm{E}[|U_j|^3|(\mathbf{Z} = \mathbf{z})] \left\| S^{-1/2}_{\sigma^2\mathbf{G},\mathbf{G}} \mathbf{G}_j \right\|_2^3.$$

Proof. Given in the Appendix.

The first two results are of a familiar form. The last result is a type of Berry-Esseen bound. Although the Berry-Esseen bound looks at first sight asymptotic, it is not, it is exact. Assumption B5 is a Lyapunov-type condition. It has the asymptotic implications that under B1-B5, so as n increases

$$\sqrt{n}\Psi_n^{-1/2}\left(\widehat{\beta} - \beta\right)|(\mathbf{Z} = \mathbf{z}) \xrightarrow{d} N(0, I_{p+1}).$$

The Berry-Esseen bound provides guidance if p increases with n (so long as $S_{\mathbf{G},\mathbf{X}}$ and $S^{-1}_{\sigma^2\mathbf{G},\mathbf{G}}$ are well behaved as p increases).

Remark 1 *Assume $S_{\mathbf{X},\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_j\mathbf{X}_j^\intercal$ is non-singular. Then Theorem 4 corresponds to the classical* $\mathrm{E}[\widehat{\beta}_{LS}|(\mathbf{Z} = \mathbf{z})] = \beta$ *and*

$$\mathrm{Var}(\widehat{\beta}_{LS}|(\mathbf{Z} = \mathbf{z})) = \frac{1}{n}\Xi_n, \quad \Xi_n = S^{-1}_{\mathbf{X},\mathbf{X}} S_{\sigma^2\mathbf{X},\mathbf{X}} S^{-1}_{\mathbf{X},\mathbf{X}},$$

where $S_{\sigma^2\mathbf{X},\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n} \sigma_j^2\mathbf{X}_j\mathbf{X}_j^\intercal$. Assume $S_{\sigma^2\mathbf{X},\mathbf{X}}$ is positive definite. The corresponding Berry-Esseen bound is

$$\begin{aligned} &\sup_{A \in \mathcal{C}_{p+1}} \left| \Pr(\sqrt{n}\left(\widehat{\beta}_{LS} - \beta\right) \in A|(\mathbf{Z} = \mathbf{z})) - \Pr(N(0, \Xi_n) \in A|(\mathbf{Z} = \mathbf{z})) \right| \\ \leq \;& c_{LS}\frac{(p+1)^{1/4}}{n^{1/2}} \left(n^{-1} \sum_{j=1}^{n} \varsigma_{LS,j} \right), \end{aligned}$$

where $\mathcal{C}_{p+1}$ denotes the set of all convex subsets of R^{p+1} and

$$\varsigma_{LS,j} = \mathrm{E}[|U_j|^3|(\mathbf{Z} = \mathbf{z})] \left\| S^{-1/2}_{\sigma^2\mathbf{X},\mathbf{X}} \mathbf{X}_j \right\|_2^3.$$

Finally,

$$\sqrt{n}\Xi_n^{-1/2}\left(\widehat{\beta}_{LS} - \beta\right)|(\mathbf{Z} = \mathbf{z}) \xrightarrow{d} N(0, I_{p+1}).$$

Example 3 *(Continuing Example 2) Recall $p = 1$ then*

$$\mathrm{E}[\widehat{\beta}_1] = \beta_1, \quad \mathrm{Var}(\widehat{\beta}_1) = \frac{\sum_{j=1}^n \frac{\left(Z_j - \widetilde{Z}\right)^2 \sigma_j^2}{1+\left(Z_j-\overline{Z}\right)^2}}{\left[\sum_{j=1}^n \frac{\left(Z_j-\widetilde{Z}\right)^2}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}\right]^2}, \quad \mathrm{E}[\widehat{\beta}_{1,LS}] = \gamma, \quad \mathrm{Var}(\widehat{\beta}_{1,LS}) = \frac{\sum_{j=1}^n \left(Z_j - \overline{Z}\right)^2 \sigma_j^2}{\left[\sum_{j=1}^n \left(Z_j - \overline{Z}\right)^2\right]^2},$$

while

$$\mathrm{E}[\widehat{\beta}_0] = \beta_0, \quad \mathrm{Var}(\widehat{\beta}_0) = \sum_{j=1}^n \sigma_j^2 \lambda_j^2, \quad \mathrm{E}[\widehat{\beta}_{0,LS}] = \beta_0, \quad \mathrm{Var}(\widehat{\beta}_{0,LS}) = \sum_{j=1}^n \sigma_j^2 n^{-2},$$

where

$$\lambda_j = w_j - \left(\widetilde{Z} - \overline{Z}\right) w_j', \quad w_j = \frac{\frac{1}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}}{\sum_{j=1}^n \frac{1}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}}, \quad w_j' = \frac{\frac{\left(Z_j-\widetilde{Z}\right)}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}}{\sum_{j=1}^n \frac{\left(Z_j-\widetilde{Z}\right)^2}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}}.$$

4.2 Unconditional inference

The corresponding result for unconditional inference is stated in Theorem 5. The proof is a straightforward application of the usual limit theory for the method of moments, thinking of the problem as a type of two step estimation problem (e.g. Newey and McFadden (1994)).

Theorem 5 *Assume $\theta = (\psi^\intercal, \beta^\intercal)^\intercal \in \Theta$, a compact parameters space. Assume that the pairs $(\mathbf{Z}_1, Y_1), ..., (\mathbf{Z}_n, Y_n)$ are i.i.d. obeying A1-A3 and write θ_0 as the true values under this sampling. Writing $U_1 = Y_1 - \mathbf{X}_1^\intercal \beta_0$, $\mathbf{X}_1 = \mathbf{Z}_1 - \psi_0$ and $\mathbf{G}_1 = \|\mathbf{X}_1\|_2^{-1} \mathbf{X}_1$, then as $n \to \infty$, so*

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{d} N(0, \left\{\mathrm{E}[\mathbf{G}_1\mathbf{X}_1^\intercal]\right\}^{-1} \mathrm{E}(\sigma_1^2 \mathbf{G}_1\mathbf{G}_1^\intercal) \left\{\mathrm{E}[\mathbf{G}_1\mathbf{X}_1^\intercal]\right\}^{-1}),$$

assuming $\mathrm{E}[\sigma_1^2] < \infty$, where $\sigma_1^2 = \mathrm{Var}(Y_1|\mathbf{Z}_1)$.

Proof. Given in the Appendix.

Section 3 showed that the existence of $\mathrm{E}[\sigma_1^2] < \infty$ and $\mathrm{E}[\mathbf{X}_1]$ is a sufficient condition for $\mathrm{E}[\mathbf{G}_1\mathbf{X}_1^\intercal]$ and $\mathrm{E}(\sigma_1^2\mathbf{G}_1\mathbf{G}_1^\intercal)$ to both exist. Hence the central limit theory for $\widehat{\beta}$ can hold in cases where the variance of Y_1 and the variance of $\mathbf{X}_1$ do not exist. What is needed is the existence of the conditional variance of the outcomes given the predictors. To remove the conditional variance assumption a switch in estimand is needed, e.g. to one based on quantiles.

Remark 2 *The result above compares to the classical*

$$\sqrt{n}\left(\widehat{\beta}_{LS} - \beta\right) \xrightarrow{d} N\left(0, \left\{\mathrm{E}[\mathbf{X}_1\mathbf{X}_1^\intercal]\right\}^{-1} \mathrm{E}[\sigma_1^2\mathbf{X}_1\mathbf{X}_1^\intercal] \left\{\mathrm{E}[\mathbf{X}_1\mathbf{X}_1^\intercal]\right\}^{-1}\right),$$

assuming $\mathrm{E}[\mathbf{X}_1\mathbf{X}_1^\intercal]$ and $\mathrm{E}[\sigma_1^2\mathbf{X}_1\mathbf{X}_1^\intercal]$ exist and $\mathrm{E}[\mathbf{X}_1\mathbf{X}_1^\intercal]$ is invertible.

Example 4 *(Continuing Example 2) When $p = 1$, then write $\widetilde{\mathrm{E}}[Z_1] = \frac{\mathrm{E}\left[\|\mathbf{X}_1\|_2^{-1} Z_1\right]}{\mathrm{E}\left[\|\mathbf{X}_1\|_2^{-1}\right]}$, recalling $\|\mathbf{X}_1\|_2 = \sqrt{1 + (Z_1 - \mathrm{E}[Z_1])^2}$, so*

$$Avar(\widehat{\beta}_1) = \frac{1}{n} \frac{\mathrm{E}\left[\sigma_1^2 \frac{(Z_1 - \widetilde{\mathrm{E}}[Z_1])^2}{1+(Z_1-\mathrm{E}[Z_1])^2}\right]}{\mathrm{E}\left[\frac{(Z_1 - \widetilde{\mathrm{E}}[Z_1])^2}{\sqrt{1+(Z_1-\mathrm{E}[Z_1])^2}}\right]^2}, \quad Avar(\widehat{\beta}_{1,LS}) = \frac{1}{n} \frac{\mathrm{E}\left[\sigma_1^2 (Z_1 - \mathrm{E}[Z_1])^2\right]}{\mathrm{E}\left[(Z_1 - \mathrm{E}[Z_1])^2\right]^2}.$$

Remark 3 *The i.i.d. assumption on the sequence of pairs $(\mathbf{Z}_1, Y_1), (\mathbf{Z}_2, Y_2), ..., (\mathbf{Z}_n, Y_n)$ in Theorem 5 is not what drives the result. That assumption can be replaced by assuming the sequence is a martingale difference with respect to the sequence's natural filtration.*

4.3 Estimating the standard errors

In practice estimating $\mathrm{E}[\sigma_1^2 \mathbf{G}_1 \mathbf{G}_1^{\mathsf{T}}]$ or $\mathrm{E}[\sigma_1^2 \mathbf{X}_1 \mathbf{X}_1^{\mathsf{T}}]$ is delicate. The estimation challenge for thick tailed predictors has been understated in the applied finance literature.

Focus on the scalar predictor case with no intercept, so $\beta = \beta_1$, to concentrate on the important ideas. Then $\mathbf{G}_j = 1$. The extension to the general case is immediate. Then

$$\widehat{U}_j = Y_j - X_j \widehat{\beta} = (Y - X_j \beta_1) - X_j \left(\widehat{\beta}_1 - \beta_1\right) = U_j - X_j \left(\widehat{\beta} - \beta\right),$$

so define

$$\begin{aligned} \widehat{\mathrm{Var}(U_1)} &= \frac{1}{n} \sum_{j=1}^{n} \widehat{U}_j^2 \\ &= \frac{1}{n} \sum_{j=1}^{n} U_j^2 + \left(\widehat{\beta}_1 - \beta_1\right)^2 \frac{1}{n} \sum_{j=1}^{n} X_j^2 - 2 \left(\widehat{\beta}_1 - \beta_1\right) \frac{1}{n} \sum_{j=1}^{n} U_j X_j, \end{aligned}$$

as an estimator of $\mathrm{Var}(U_1)$. Then, $\frac{1}{n} \sum_{j=1}^{n} U_j^2$ converges to $\mathrm{Var}(U_1)$ using the strong law of large numbers as $\mathrm{E}[U_1] = 0$. What happens to the two other terms in the above expression?

Now, if $\mathrm{E}[X_1^2] < \infty$, under the conditions of Theorem 5,

$$\left\{\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right)\right\}^2 \frac{1}{n} \sum_{j=1}^{n} X_j^2 \xrightarrow{d} \mathrm{E}X_1^2 \frac{\mathrm{Var}(U_1)}{\{\mathrm{E}|X_1|\}^2} \chi_1^2,$$

while

$$\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) \left(\frac{1}{n} \sum_{j=1}^{n} U_j X_j\right) \xrightarrow{p} 0,$$

as long as $\mathrm{E}[U_1 X_1]$ exists (in which case $\mathrm{E}[U_1 X_1] = 0$). Then, under the conditions of Theorem 5,

$$\widehat{\mathrm{Var}(U_1)} \xrightarrow{p} \mathrm{E}[\sigma^2(X_1)] = \mathrm{Var}(U_1).$$

But what happens if $\mathrm{E}[X_1^2]$ does not exists? The CLT for $\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right)$ does not change and $\frac{1}{n}\sum_{j=1}^n U_j^2$ is well behaved. However, trouble brews in the terms $\frac{1}{n}\sum_{j=1}^n X_j^2$ and $\frac{1}{n}\sum_{j=1}^n U_j X_j$. In our simulations, $\widehat{\mathrm{Var}(U_1)}$ becomes inadequate if $\mathrm{E}[X_1^2]$ ceases to exist.

To entirely circumvent this problems, I use a weighted estimator, where the weights will be denoted $w(x)$,

$$\begin{aligned}
\widetilde{\mathrm{Var}(U_1)} &= \frac{1}{n}\sum_{j=1}^n \left\{\widehat{U}_j w(X_j)\right\}^2, \quad w(x) = 1_{\frac{|x|}{\mathrm{E}|X_1|} < dn^{1/5}} \\
&= \frac{1}{n}\sum_{j=1}^n U_j^2 w(X_j) + \left(\widehat{\beta}_1 - \beta_1\right)^2 \frac{1}{n}\sum_{j=1}^n X_j^2 w(X_j) \\
&\quad -2\left(\widehat{\beta}_1 - \beta_1\right)\frac{1}{n}\sum_{j=1}^n U_j X_j w(X_j).
\end{aligned}$$

This clips regression residuals associated with large predictors. Then,

$$\begin{array}{lcl}
\frac{1}{n}\sum_{j=1}^n U_j^2 1_{|X_j| < dn^{1/5}\mathrm{E}|X_1|} & \xrightarrow{p} & \mathrm{E}[\sigma^2(X_1)] \\
\frac{1}{n}\sum_{j=1}^n X_j^2 1_{\frac{|x_j|}{\mathrm{E}|X_1|} < dn^{1/5}} & \leq & d^2 n^{2/5}\left(\mathrm{E}|X_1|\right)^2, \\
\left|\frac{1}{n}\sum_{j=1}^n U_j X_j w(X_j)\right| & \leq & dn^{1/5}\mathrm{E}|X_1|\frac{1}{n}\sum_{j=1}^n |U_j|.
\end{array}$$

As $\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) = O_p(1)$, these three results imply that

$$\widetilde{\mathrm{Var}(U_1)} \xrightarrow{p} \mathrm{E}[\sigma^2(X_1)],$$

even if $\mathrm{E}[X_1^2]$ does not exist.

In our simulation and empirical work we take $d = 10$. If $n = 100$, the scaled threshold is $dn^{1/5} \simeq 16$. This implies the weight function will not clip any regression residual unless the associated predictor is extraordinarily unusual. For thin tailed predictors, clipping is neither helpful or harmful. For thick tailed predictors, it is deeply important. As researchers usually do not know the tail behavior of their predictors, the safe approach is to always include the weighting function.

More broadly, the same line of argument applies to

$$\frac{1}{n}\sum_{j=1}^n \left\{\widehat{U}_j w(\mathbf{X}_j)\right\}^2 \mathbf{G}_j \mathbf{G}_j^{\intercal} \xrightarrow{p} \mathrm{E}(\sigma_1^2 \mathbf{G}_1 \mathbf{G}_1^{\intercal}),$$

where now $w(\mathbf{x}) = 1_{\left\{\max_i \frac{|x_i|}{\mathrm{E}|X_{1,i}|}\right\} < dn^{1/5}}$, so long as $\mathrm{E}[\sigma_1^2] < \infty$. More straightforwardly, by the strong law of large numbers

$$\frac{1}{n}\sum_{j=1}^n \mathbf{G}_j \mathbf{X}_j^{\intercal} \xrightarrow{p} \mathrm{E}(\mathbf{G}_1 \mathbf{X}_1^{\intercal}),$$

so long as $\mathrm{E}(\mathbf{G}_1 \mathbf{X}_1^{\intercal})$ exists – which is true so long as $\mathrm{E}|\mathbf{X}_1|$ exists.

Thus asymptotically valid estimators of the standard errors can be computed without needing any more assumptions about higher order moments.

As mentioned in the introduction, for least squares, their robust standard errors need at least the fourth moments of the predictors to be valid. This is spelt out in Assumption 4 of White (1980).

Example 5 *(Continuing from Example 2) When $p = 1$ then*

$$\widehat{\mathrm{Var}(\widehat{\beta}_1)} = \frac{\sum_{j=1}^n W_j \frac{\left(Z_j-\widetilde{Z}\right)^2 \widehat{U}_j^2}{1+\left(Z_j-\overline{Z}\right)^2}}{\left[\sum_{j=1}^n \frac{\left(Z_j-\widetilde{Z}\right)^2}{\sqrt{1+\left(Z_j-\overline{Z}\right)^2}}\right]^2}, \quad \widehat{\mathrm{Var}(\widehat{\beta}_{1,LS})} = \frac{\sum_{j=1}^n \left(Z_j-\overline{Z}\right)^2 \widehat{U}_{j,LS}^2}{\left[\sum_{j=1}^n \left(Z_j-\overline{Z}\right)^2\right]^2},$$

where $\widehat{U}_j = \left(Y_j - \widetilde{Y}\right) - \widehat{\gamma}\left(X_j - \widetilde{X}\right)$, $\widehat{U}_{j,LS} = \left(Y_j - \overline{Y}\right) - \widehat{\gamma}_{LS}\left(X_j - \overline{X}\right)$ *and* $W_j = 1_{\frac{|X_i-\overline{X}|}{|X-\overline{X}|} < dn^{1/5}}$.

5 Simulation experiment

In this section I focus on the performance of the approximate pivots

$$\widehat{T}_{\widehat{\beta}} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\widehat{\mathrm{Var}(\widehat{\beta}_1)}}}, \quad \text{and} \quad \widehat{T}_{LS} = \frac{\widehat{\beta}_{1,LS} - \beta_1}{\sqrt{\widehat{\mathrm{Var}(\widehat{\beta}_{1,LS})}}}$$

in the case with an intercept and one predictor, so $\beta = (\beta_0, \beta_1)^T$, where the weight function for $\widehat{T}_{\widehat{\beta}}$ is $w(x) = 1_{|x|<\widehat{c}10n^{0.2}}$, $\widehat{c} = \frac{1}{n}\sum_{j=1}^n |X_j - \overline{X}|$. The truncation in $w(x)$ has no impact except for the very extreme cases we discuss in Section 5.3, which studies $\widehat{T}_{\widehat{\beta}}$ in the case where $\mathrm{Var}(X_1) = \infty$. Recall, theory suggests that weighting is needed in that case.

The simulation design I initially use is based around regressing stock returns on a broad based index, to estimate a "beta". The design mimics the empirical challenge tackled in the next section. That challenge looks at 2 years of weekly percentage arithmetic returns on a major U.S. company, Y_j, and X_j will be the S&P500 index arithmetic returns. In the empirical work this will be implemented for more than 400 individual companies, using hypothesis tests based on $\widehat{T}_{\widehat{\beta}}$ and $\widehat{T}_{LS}$ to identify stocks with very high or very low betas.

5.1 Initial experiment

Assume

$$X_j \overset{indep}{\sim} \psi + \sigma_X \sqrt{\frac{\nu-2}{\nu}} V_j, \quad V_1 \sim t_\nu, \quad \nu > 2, \quad j = 1, ..., n,$$

and

$$Y_j | \left(X_j = x\right) \overset{indep}{\sim} N(\beta_0 + \beta_1(x - \psi), \sigma^2).$$

This implies that, for every value of ν, the $\mathrm{E}[X_1] = \psi$, $\mathrm{E}[Y_1] = \beta_0$, and $\mathrm{Var}(X_1) = \sigma_X^2 < \infty$. The simulations will have homoskedasticity, but the approximate pivots will be computed without imposing that. Importantly if $\nu < 4$ the $\widehat{T}_{LS}$ is not asymptotically N(0,1), while $\widehat{T}_{\widehat{\beta}}$ will be.

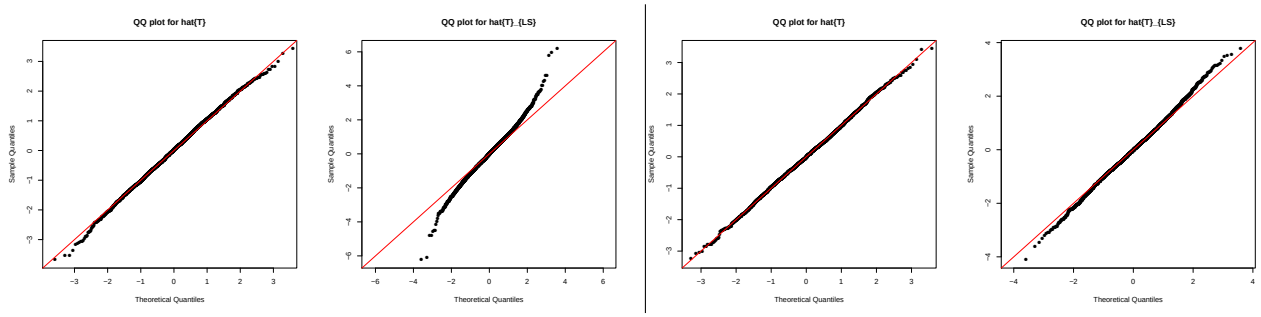

Figure 2: QQ plots for $\widehat{T}_{\widehat{\beta}}$ and $\widehat{T}_{LS}$. Simulation designed to match empirical data we see for weekly financial returns on U.S. stocks for major companies, where the predictor is the main major index. The left hand side corresponds to $\nu = 2.4$, the right hand side has $\nu = 4.4$. Throughout, $n = 100$.

Hence $\widehat{T}_{LS}$ is expected to have poor performance unless ν is substantially above 4. This is what you will see in the simulations.

To calibrate this using universally available data, I look at weekly arithmetic (total) returns on the SPDR S&P 500 ETF Trust (SPY) from 1st August 2018 to 4th August 2020, downloaded using the `R` package `Quantmod` from Yahoo's database. The `R` code for this download, delivering a vector of weekly returns "X" is

```
getSymbols("SPY",from='2018-08-01',to='2020-08-04',verbose = TRUE); head(SPY);
XdataM = (to.weekly(SPY))$SPY.Adjusted
X = data.matrix(100*diff(XdataM)/lag(XdataM))[-1];
```

The sample mean and standard deviation suggest taking $\sigma_X = 3.24$, $\psi = 0.21$, and the ML estimator of ν, computed using `R`'s function `fitdistr` for the student-t distribution, is 2.16 with a standard error of around 0.5. This is not an unusual result — typical extreme value theory estimates of the tail index of equity indexes suggest 2 but not 4 moments exist.

To nudge towards safer grounds for least squares, in the simulations we will use $\nu = 2.4$, a little larger than I saw in the data. Later, we will explore many different values of ν.

The initial focus is on the case where $n = 100$, $\psi = 0.21$, $\beta_0 = 0$, $\nu = 2.4$, $\sigma = 2$, $\beta_1 = 1$ and $\sigma_X = 3.24$. In this case, the predictors have 2 but not 3 moments.

I will initially summarize results using QQ plots, based on 3,000 replications, comparing the simulated quantiles of $\widehat{T}_{\widehat{\beta}}$ and $\widehat{T}_{LS}$ to quantiles of their $N(0,1)$ baseline.

The resulting QQ plots for $\widehat{T}_{\widehat{\beta}}$ and $\widehat{T}_{LS}$ are given in the left hand side of Figure 2 for $\nu = 2.4$. The results for $\widehat{T}_{\widehat{\beta}}$ are strong, matching the Gaussian limit theorem throughout except perhaps in the very extreme tails. The results for $\widehat{T}_{LS}$ are terrible. Perhaps poor behavior for $\widehat{T}_{LS}$ is to be expected given ν is low.

The right hand side of Figure 2 gives the results for the corresponding easier case of $\nu = 4.4$. The performance of $\widehat{T}_{\widehat{\beta}}$ does not change very much, again providing very solid results. The $\widehat{T}_{LS}$ is much better than in the heavier tail case, no longer terrible, just poor. This $\nu = 4.4$ case is a situation where the limit theory for $\widehat{T}_{LS}$ is valid, it is just not very accurate in practice.

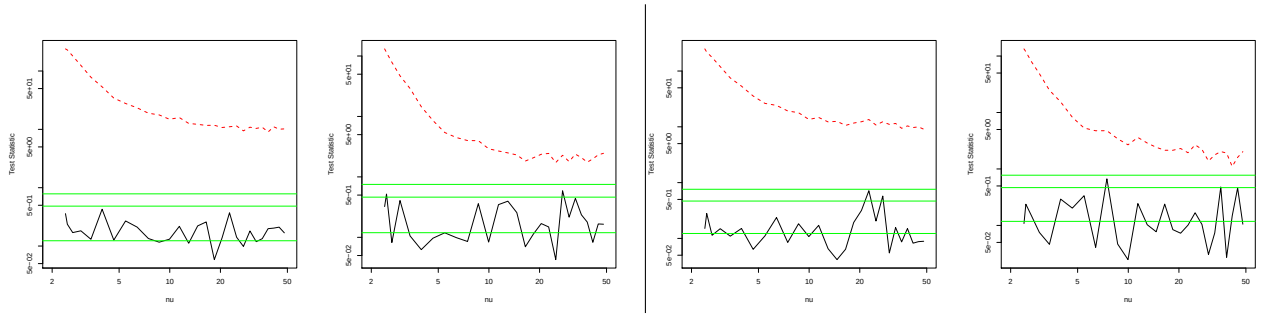

Figure 3: Cramér-von-Mises statistics test of N(0,1) for $\widehat{T}_{\widehat{\beta}}$ (black line) and $\widehat{T}_{LS}$ (dotted red line) for various values of n, ν and σ. Throughout 500,000 replications are used. Green horizontal lines represent the 0.5, 0.95 and 0.99 quantiles of the Cramer-Von-Mises statistic when the data is i.i.d. $N(0,1)$. Results to the left hand side have $\sigma = 2$, and results to the right have $\sigma = 4$. 1st and 3rd graph have $n = 100$; 2nd and 4th have $n = 250$. Throughout the x-axis is ν, on the log-scale. The y-axis is also drawn on the log-scale.

5.2 More extensive experiments

To compare performance in a wide set of diverse environments, I used the Cramér-von-Mises statistic to measure how non-Gaussian 500,000 replications of $\widehat{T}_{\widehat{\beta}}$ and $\widehat{T}_{LS}$ were, for a variety of values of n and ν, as well as σ. The Cramér-von-Mises test statistic is reviewed in Baringhouse and Henze (2017). It is viewed as one of the most powerful distributional tests. To benchmark the values of the Cramér-von-Mises statistic of normality, the green horizontal lines are the 0.5, 0.95 and 0.99 quantiles of the distribution of the Cramér-von-Mises test statistic computed using 500,000 replications under the null of i.i.d. $N(0,1)$. The test is implemented in `R` using the function `cvm.test(data, "pnorm",mean=0,sd=1)`.

The results are given in Figure 3. Crucially, notice that all plots use log-scales on both the x-axis and the y-axis. The dotted red line is the result for $\widehat{T}_{LS}$, while the black line is the corresponding results for $\widehat{T}_{\widehat{\beta}}$. The x-axis is the value of ν; the y-axis is the Cramér-von-Mises test statistic. The 1st and 3rd graph have $n = 100$, the 2nd and 4th have $n = 250$. The left hand side corresponds to $\sigma = 2$, the right hand side $\sigma = 4$.

The results for $\widehat{T}_{\widehat{\beta}}$ are encouraging. Even with 500,000 replications it is typically not possible to reasonably reject normality of $\widehat{T}_{\widehat{\beta}}$ even with $n = 100$. When ν is at the bottom end of the plots, $\nu = 2.4$ and $n = 100$, there is some evidence of a tiny amount of non-normality. The value of σ does not impact the results materially. As n increases to 250 the results improve, a little.

For $\widehat{T}_{LS}$ the results uniformly reject normality, typically dramatically. As ν increases the rejections become less significant, as expected. As n increases to 250 the results improve, but only by a little.

5.3 Pushing to the case where $\mathrm{Var}(X_1) = \infty$

To assess the $N(0,1)$ approximation for $\widehat{T}_{\widehat{\beta}}$ when $\mathrm{Var}(X_1) = \infty$, I ran a separate experiment. This experiment is less relevant to major equity data, although some commodity market price moves have extraordinarily thick tails and it is interesting theoretically. I excluded $\widehat{T}_{LS}$ from

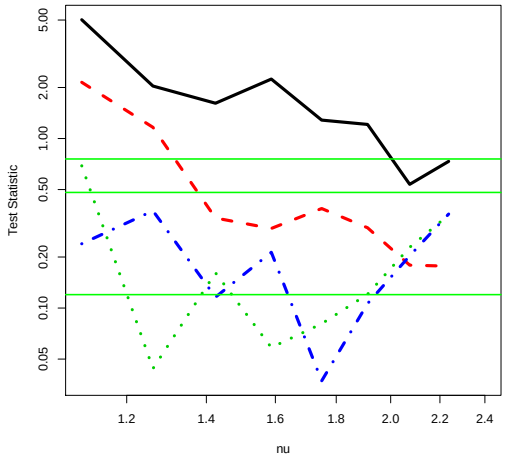

Figure 4: Cramér-von-Mises statistics test of $N(0,1)$ for $\widehat{T}_{\widehat{\beta}}$ and various values of n and σ. 500,000 replications are used. Green horizontal lines represent the 0.5, 0.95 and 0.99 quantiles of the Cramer-Von-Mises statistic when the data is i.i.d. $N(0,1)$. Throughout the x-axis is ν. The y-axis is drawn on the log-scale. Sample size are: $n = 100$ (black line), 250 (red dotted line), $1,000$ (green dots) and $10,000$ (blue line).

consideration, as the previous experiment has shown it would have weak performance and the asymptotics is far from being valid.

It was less than clear to me how to generate predictors which are broadly comparable in scale across difference values of ν. I eventually settled on

$$X_j \overset{indep}{\sim} 0.21 + \sigma_X V_1 / \mathrm{E}\,|V_1|\,, \quad V_1 \sim t_\nu$$

which scales the student-t random variable so $\mathrm{E}|X_1 - 0.21| = \sigma_X$ whatever the value of ν.

Figure 4 shows the results for $n = 100$ (black line), 250 (red dotted line), $1,000$ (green dots) and $10,000$ (blue line), here with $\sigma = 2$ throughout. The Figures again plot the Cramér-von-Mises test statistic against ν based on $500,000$ replications. The results are encouraging, although not entirely positive. For small n and small ν, significant distortion is present. Samples of around 1,000 do deliver results for $\widehat{T}_{\widehat{\beta}}$ which are hard to reject from a null of $N(0,1)$, even in cases where the predictors are just slightly less thick tailed than Cauchy.

Without the weighting function in $\widehat{T}_{\widehat{\beta}}$, these simulations fall apart when $\mathrm{Var}(X_1) = \infty$. In that situation, I found no sign that increasing n improves the behavior of $\widehat{T}_{\widehat{\beta}}$. This is in line with the suggestions from the theory than once moments somewhat above the second do to not exist then the weight function becomes useful and eventually essential if $\mathrm{Var}(X_1) = \infty$. I recommend always including the weight function in practice. It is harmless for thin tailed data and essential for very thick tailed data.

6 Empirical work

6.1 Background

Some investors, such as young workers saving into pensions who do not have a mortgage, are unable to take on the level of financial leverage they may rationally desire due to administrative rules or the inability to borrow against their human capital. One viable investment strategy is to overweight their portfolio with high beta stocks, that is, stocks which move more strongly with the main market indexes than most stocks. This is discussed in Black (1972), amongst others. In finance betas are usually measured by regressing the returns on the individual stock on a wide market index. Such betas are used directly in vast numbers of empirical papers and drive other methods such as "Fama-MacBeth regressions". All of these empirical results are fragile due to the thick tailed index returns.

The fragility of least squares estimators is known in the applied finance literature — a frequent response is to employ shrinkage, e.g. Hollstein et al. (2019), Vasicek (1973) and Karolyi (1992). The shrinkage approach typically needs an estimate of the variance of the least squares estimator for the s-th stock, $\widehat{\beta}_{LS,s}$, shrinking towards a cross-sectional average of betas, β_μ, traditionally yielding

$$\overline{\beta}_{LS,s} = \widehat{\lambda}_{LS,s}\widehat{\beta}_{LS,s} + (1-\widehat{\lambda}_{LS,s})\beta_\mu, \quad \widehat{\lambda}_{LS,s} = \frac{\sigma_\beta^2}{\widehat{\mathrm{Var}(\widehat{\beta}_{LS,s})} + \sigma_\beta^2}, \quad s = 1, 2, ...,$$

where σ_β^2 is a cross-sectional variance of betas. But for thick tailed data, the $\widehat{\mathrm{Var}(\widehat{\beta}_{LS,s})}$ are unreliable, which pollutes $\overline{\beta}_{LS,s}$. Shrinkage based on $\widehat{\beta}_s$, the beta estimator for the s-th stock, would yield

$$\overline{\beta}_s = \widehat{\lambda}_s\widehat{\beta}_s + (1-\widehat{\lambda}_s)\beta_\mu, \quad \widehat{\lambda}_s = \frac{\sigma_\beta^2}{\widehat{\mathrm{Var}(\widehat{\beta}_s)} + \sigma_\beta^2}, \quad s = 1, 2, ...,$$

which should not be derailed by $\widehat{\mathrm{Var}(\widehat{\beta}_s)}$. I will not report $\overline{\beta}_s$ empirically in this paper.

Less attention has been paid in the finance literature to the fragility of associated inference — which is my interest here.

Typically high beta investments will have high risk in order to potentially capture high expected returns. The opposite of this is attractive to a different type of investor. Some investors search out low beta stocks, hoping to have low risk, positive risk premium although relatively low expected return. This is discussed in Baker et al. (2011), who also review the extensive literature on this topic.

6.2 Selection by hypothesis test

But how to select high beta stocks and low beta stocks? Once selected, these groups of stocks could potentially be placed into portfolios or packaged as low and high beta ETFs.

In this Section selection will be regarded as a hypothesis testing problem. The s-th stock will be labelled a high beta stock if we can reject the null

$$H_0 : \beta_s \leq 1.4, \quad \text{against} \quad H_1 : \beta_s > 1.4,$$

where β_s is the beta of the s-th stock.

I will label the s-th stock a low beta stock if the null

$$H_0 : \beta_s \geq 0.8, \quad \text{against} \quad H_1 : \beta_s < 0.8,$$

is rejected.

The tests will be based on the approximate pivots

$$\widehat{T}_{\widehat{\beta}} \quad \text{and} \quad \widehat{T}_{LS}$$

rejecting the nulls using a one-sided test with nominal size of 5%. I will compare the results for $\widehat{T}_{\widehat{\beta}}$ with the one based on $\widehat{T}_{LS}$.

6.3 Data

I downloaded the data from Yahoo for stock prices from 1 August 2018 to 4th August 2020. The S&P500 was measured using the SPDR S&P 500 ETF Trust (SPY). I converted these into weekly percentage arithmetic returns. These weekly returns will be compared to the returns on 416 individual stocks, which are components of the S&P500. The list of the stocks is available in `RegFinance1.r`, which produces all the results given in this Section.

Why weekly returns? There are virtues in using higher frequency data than weekly returns to estimate betas. In theory they can produce vastly more precise estimates. High frequency versions of regression include Barndorff-Nielsen and Shephard (2004), Barndorff-Nielsen et al. (2011) and Bollerslev et al. (2020). These sophisticated data hungry methods try to overcome the impact of nonsynchronous trading and differential rates of price discovery. However, they miss the impact of overnight returns. Daily returns capture overnight effects, but have significant lead-lag correlations. The hope with weekly returns is that most of the impact of these dependencies will be averaged away or dwarfed by other long-term effects. Many practitioners go further than this and use 2 to 5 years of monthly returns, but we do not follow that route. Further, some of the volatility clustering seen in finance is taken out by using weekly returns rather than high frequency returns.

6.4 Cross-sectional results

The right hand side of Figure 5 plots the cross-section of $\hat{\beta}_{LS}$ against $SE(\hat{\beta}_{LS})$. The left hand side gives the corresponding result for $\hat{\beta}$. The major impact of moving from $\hat{\beta}_{LS}$ to $\hat{\beta}$ is that $\hat{\beta}_{LS}$ delivers some estimators with tiny standard errors, which is not the case with $\hat{\beta}$. Indeed the smallest standard error for $\hat{\beta}_{LS}$ in the cross-section is around 0.054, while the minimum for $SE(\hat{\beta})$ is around 0.080 — around 50% higher.

Table 1 provides summary statistics on the cross-section of $\hat{\beta}$ and $\hat{\beta}_{LS}$. The estimates have roughly the same level, with roughly the same spread. The $\hat{\beta}$ are slightly lower than the $\hat{\beta}_{LS}$ in this sample, across the distribution. Over the entire cross-section $SE(\hat{\beta})$ is a little above $SE(\hat{\beta}_{LS})$, as we would expect, but the difference is very modest. This suggests that a potential worry over $\hat{\beta}$ being generally much less precise than $\hat{\beta}_{LS}$ is not compelling here. Table 1 also details the cross-section of $\widehat{\alpha}$ against $\widehat{\alpha}_{LS}$. These are very similar, which is also true of the $SE(\widehat{\alpha})$ and $SE(\widehat{\alpha}_{LS})$.

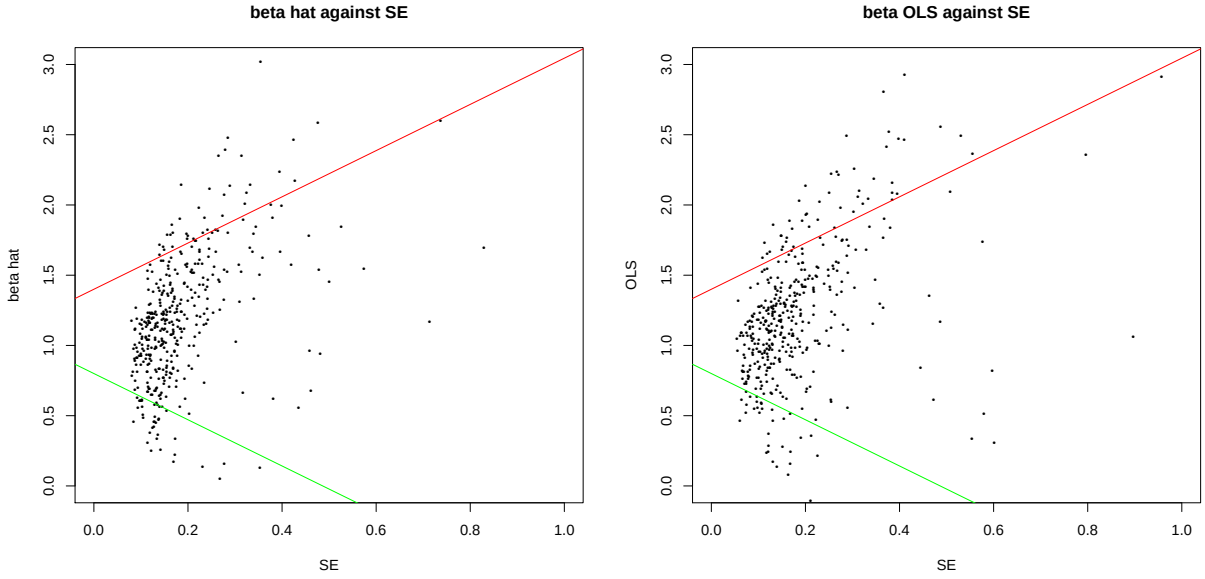

Figure 5: The left hand side is the cross-section of $\widehat{\beta}$ against its standard error. The right hand side shows the corresponding results for $\widehat{\beta}_{LS}$. The red line has an intercept of 1.4 and slope of 1.64. Values above the red line are labelled high beta stocks. The green line has an intercept of 0.8 and slope of -1.64. Values below the green line are labelled low beta stocks.

	$\hat{\beta}$	$\hat{\beta}_{LS}$	$SE(\hat{\beta})$	$SE(\hat{\beta}_{LS})$	$\hat{\alpha}$	$\hat{\alpha}_{LS}$	$SE(\hat{\alpha})$	$SE(\hat{\alpha}_{LS})$
Mean	1.17	1.21	0.184	0.184	0.067	0.100	0.402	0.420
Q(0.1)	0.619	0.644	0.103	0.087	-0.548	-0.427	0.240	0.247
Q(0.5)	1.16	1.17	0.155	0.152	0.164	0.148	0.347	0.359
Q(0.9)	1.76	1.83	0.287	0.304	0.546	0.547	0.603	0.663

Table 1: Cross-sectional summaries of the betas and alphas, estimated by $\hat{\beta}$, $\hat{\beta}_{LS}$ and $\hat{\alpha}$, $\hat{\alpha}_{LS}$. The cross-section is over 400 individual stock returns, individually regressed against the S&P500 index.

The left hand side of Figure 6 plots $\hat{\beta}$ against $\hat{\beta}_{LS}$ in the cross-section. The blue line is a 45 degree line, while a cross-sectional regression of $\hat{\beta}$ against $\hat{\beta}_{LS}$ yields an intercept of 0.150 (S.E. of 0.026) and slope of 0.849. The $R^2 \simeq 0.816$. The least squares linear regression line is shown in the figure by the red line. Overall the picture shows the two sets of estimates are comparable. The $\hat{\beta}$ tend to be slightly pulled up for low betas and pulled down for high betas, when lined up with the corresponding $\hat{\beta}_{LS}$.

The right hand side of Figure 6 plots $SE(\hat{\beta})$ against $SE(\hat{\beta}_{LS})$ in the cross-section. The blue line is a 45 degree line, while a cross-sectional regression of $SE(\hat{\beta})$ against $SE(\hat{\beta}_{LS})$ yields an intercept of 0.045 (S.E. of 0.004) and slope of 0.757. The $R^2 \simeq 0.806$. The least squares linear regression line is shown in the Figure by the red line. When the S.E.s are low the $SE(\hat{\beta})$ is materially higher than the $SE(\hat{\beta}_{LS})$. However, for high S.E.s this is not the case. There is more discordance between $SE(\hat{\beta})$ and $SE(\hat{\beta}_{LS})$ than $\hat{\beta}$ and $\hat{\beta}_{LS}$. In terms of $SE(\widehat{\alpha})$ against $SE(\widehat{\alpha}_{LS})$, they have a correlation of about 0.94, so are not plotted here.

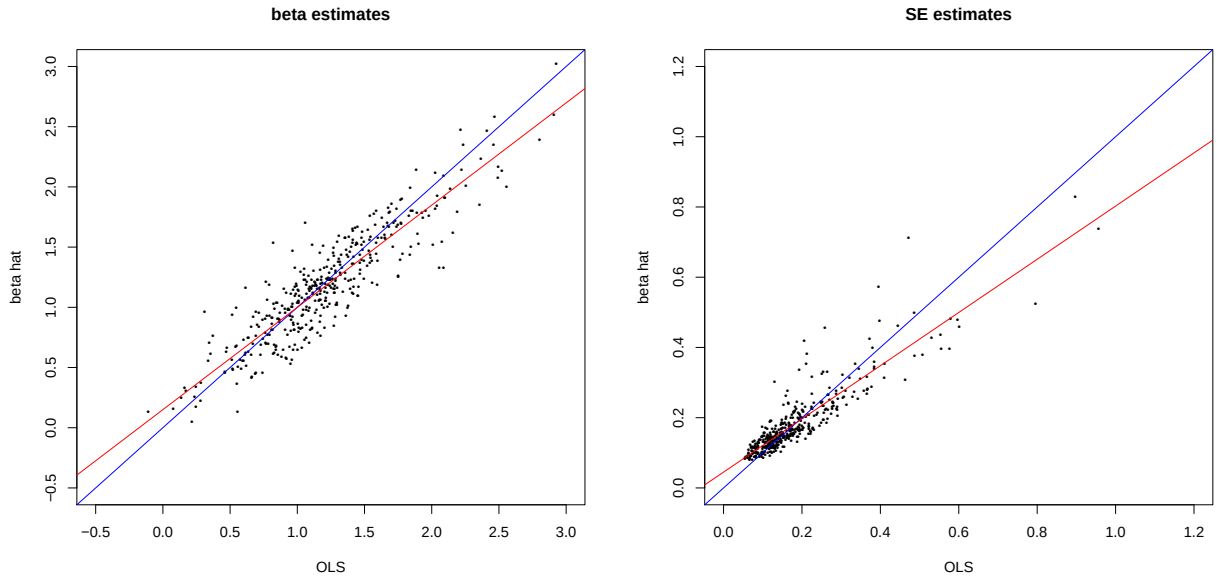

Figure 6: The left hand side is a plot of estimated beta for over 400 different stocks based on weekly arithmetic returns, regressing the returns against the S&P500 market index. The y-axis is $\hat{\beta}$, the x-axis is $\hat{\beta}_{LS}$. The blue line is a 45 degree line, going through the origin. The red line is a least squares fitted straightline from regressing $\hat{\beta}$ on $\hat{\beta}_{LS}$. The right hand side is the corresponding estimated standard errors for $\hat{\beta}$ on $\hat{\beta}_{LS}$.

Overall, these summary measures suggest that inference based on $\hat{\beta}$ and $SE(\hat{\beta})$ may yield less extreme conclusions than those based on $\hat{\beta}_{LS}$ and $SE(\hat{\beta}_{LS})$.

Figure 7 mimics Figure 5, but now for alpha. The differences between $\widehat{\alpha}$,$SE(\widehat{\alpha})$ and $\widehat{\alpha}_{LS}$,$SE(\widehat{\alpha}_{LS})$ are much smaller than in the beta case. Inference based on $\widehat{\alpha}$ and $SE(\widehat{\alpha})$ should be very similar to that based on $\widehat{\alpha}_{LS}$ and $SE(\widehat{\alpha}_{LS})$.

Figure 5 shows the critical values for $\hat{\beta}$ and $\hat{\beta}_{LS}$ for the high beta null, plotted as the red line (with an intercept of 1.4 and slope of 1.64). Any estimate above the red line corresponds to a statistically significant high beta. The corresponding results below the green line are statistically significant low beta stocks.

Table 2 provides a summary of the results from the tests for the nulls for the high and low betas. For both tests, it gives the number of rejections of the null together with the number of agreements. For the high beta stocks $\widehat{T}_{LS}$ tests are much more liberal than $\widehat{T}_{\widehat{\beta}}$. It is rare for $\widehat{T}_{\widehat{\beta}}$ to find a high beta stock which is not found by $\widehat{T}_{LS}$, but common to see high beta stocks selected by $\widehat{T}_{LS}$ but not $\widehat{T}_{\widehat{\beta}}$. The results are more scattered for the test of the low beta stocks.

This is highlighted by Figure 8, plotting $\hat{\beta}$ against $\hat{\beta}_{LS}$ for selected stocks. The 1st and 3rd selection is carried out by $\widehat{T}_{\widehat{\beta}}$, the 2nd and 4th by $\widehat{T}_{LS}$. The left hand side of the Figure shows results for the selection of high beta stocks. The right hand side gives the corresponding selected low beta stocks. The labels are the individual stock tickers.

For the high beta selections, the $\widehat{T}_{\widehat{\beta}}$ selected stocks have both estimators $\hat{\beta}$ and $\hat{\beta}_{LS}$ indicating pretty high betas. When the selection is based on $\widehat{T}_{LS}$ the results are more variable.

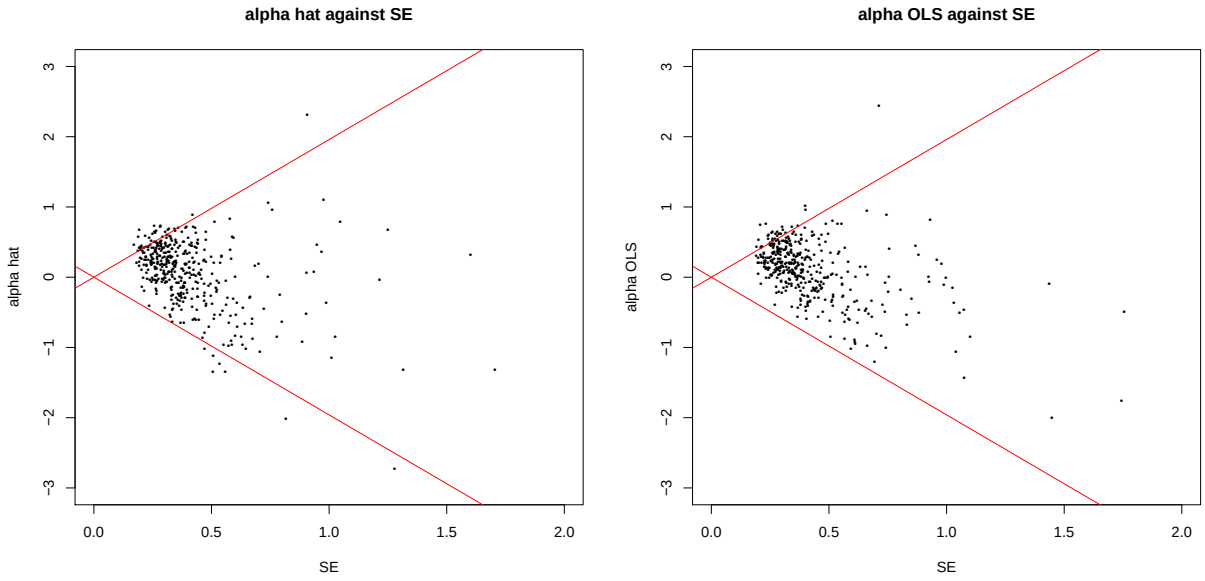

Figure 7: The right hand side is the cross-section of least squares estimates $\widehat{\alpha}_{LS}$ against the estimated standard error. The left hand side has the same object for $\widehat{\alpha}$ and its standard error. The red lines have an intercept of 0 and slopes of ± 1.96. Values above or below the red line are labelled significant alpha stocks.

	$\widehat{T}_{\widehat{\beta}}$	$\widehat{T}_{LS}$	Agree
High beta	36	49	32
Low beta	37	32	23

Table 2: Number of rejections of the null. At a nominal 5% level, the procedure would expect around 20 rejections if the null holds just by chance. The cross-section is over 400 individual stock returns, individually regressed against the S&P500 index.

For low beta selection, the story is much more mixed.

6.5 Rolling betas

Empirical researchers often deal with time-varying betas by using moving block, or rolling, averages — see Engle (2016) for an alternative model based approach and a discussion of the literature. In our context a rolling average approach computes the statistics $\left(\hat{\beta}, \hat{\beta}_{LS}\right)$ and $(SE(\hat{\beta}), SE(\hat{\beta}_{LS}))$ on the last 100, say, weeks of data, moving that window through time.

What do the pairs $(\hat{\beta},\hat{\beta}_{LS})$, $(SE(\hat{\beta}), SE(\hat{\beta}_{LS}))$ and the two pair of pairs $(\widehat{T}_{\widehat{\beta}}, \widehat{T}_{LS})$ (corresponding to the high and low beta hypotheses) look like over the last 15 years for the first stock in our database, ABT, Abbott Laboratories? Now the data starts on 1 January 2005 and the rolling window always covers 100 weeks of data. In this entire database there are 358 stocks.

Figure 9 contains the results for ABT: $\hat{\beta}$ is drawn using a thick black line; $\hat{\beta}_{LS}$ using a

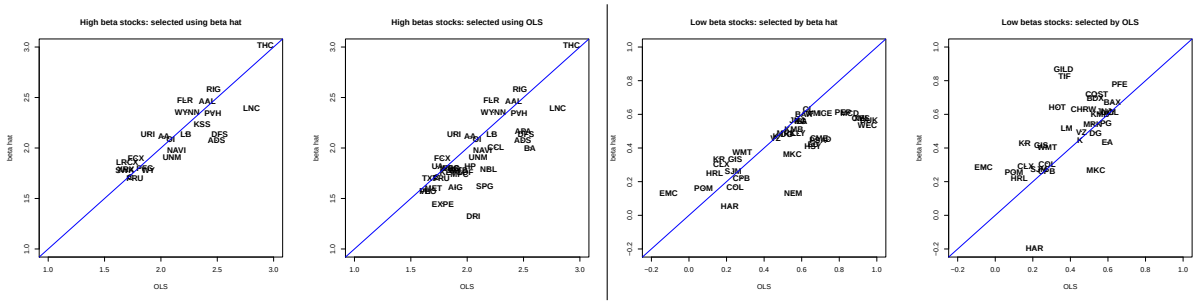

Figure 8: LHS shows selection of large beta stocks, RHS has selection of small beta stocks. The selection in the 1st and 3rd graph is carried out using $\widehat{T}_{\widehat{\beta}}$, the 2nd and 4th are implemented using $\widehat{T}_{LS}$. The tickers are used to plot individual stock selections.

$\sqrt{\mathrm{E}[(100\Delta\widehat{\beta})^2]}$	$\sqrt{\mathrm{E}[(100\Delta\widehat{\beta}_{LS})^2]}$	$\sqrt{\mathrm{E}[(100\Delta SE(\widehat{\beta}))^2]}$	$\sqrt{\mathrm{E}[(100\Delta SE(\widehat{\beta}_{LS}))^2]}$
2.44	2.84	0.152	0.269

Table 3: Measures of the roughness of the path of estimated beta and corresponding estimated standard errors. The roughness is measured by the square root of the average squared 100 times daily changes in the estimator. Here Δ is a time series difference operator.

dotted red line. The beta estimates ($\hat{\beta}$,$\hat{\beta}_{LS}$) broadly track one another. The $\hat{\beta}_{LS}$ coefficient does not go as high as the $\hat{\beta}$ during higher beta periods, nor as low during lower beta periods.

I would like you to mostly focus on is the right hand side picture, which plots the pair $(SE(\hat{\beta}), SE(\hat{\beta}_{LS}))$ through time. Although they follow the same general level through time, $SE(\hat{\beta}_{LS})$ is very rough, sometimes moving dramatically over a few datapoints, as individual pairs of data fall in or out of the 100 day window. This is exactly what was feared in the introduction, it is very sensitive to large moves in the predictors. $SE(\hat{\beta})$ is much smoother, as if it has been time-series smoother — but it has not been. It drifts down and then up, the range moving by a factor of 2 in the picture.

This result is typical in the cross-section. Averaging over all time periods and across all stocks, the first element of Table 3 shows the square root of the average squared 100 times the daily time series changes in $\widehat{\beta}$. This is just below 2.5, while the corresponding result for $\widehat{\beta}_{LS}$ is around 15% higher. There are much bigger differences in the roughness of the standard deviations. The average movement of $SE(\widehat{\beta}_{LS})$ is around 70% higher than the result for $SE(\widehat{\beta})$.

One implication of this can be seen in Figure 10 for ABT. On the left hand side the pair $(\widehat{T}_{\widehat{\beta}}, \widehat{T}_{LS})$ is plotted against time for the large beta hypothesis test. The test based on least squares $\widehat{T}_{LS}$ is very jagged through time, moving around week by week dramatically. The main driver of these moves are the jagged standard errors. The thick black line, corresponding to $\hat{\beta}$ is again smooth. The same holds for the small beta test picture, which is the right hand side picture. Here the test rejects the null and selects this stock as a small beta stock for roughly the same period, but this is lucky for the evidence is very strong which makes choosing between using $\widehat{T}_{\widehat{\beta}}$ and $\widehat{T}_{LS}$ moot.

The evidence from looking at rolling betas is that the classic $\widehat{\beta}_{LS}$ standard errors move

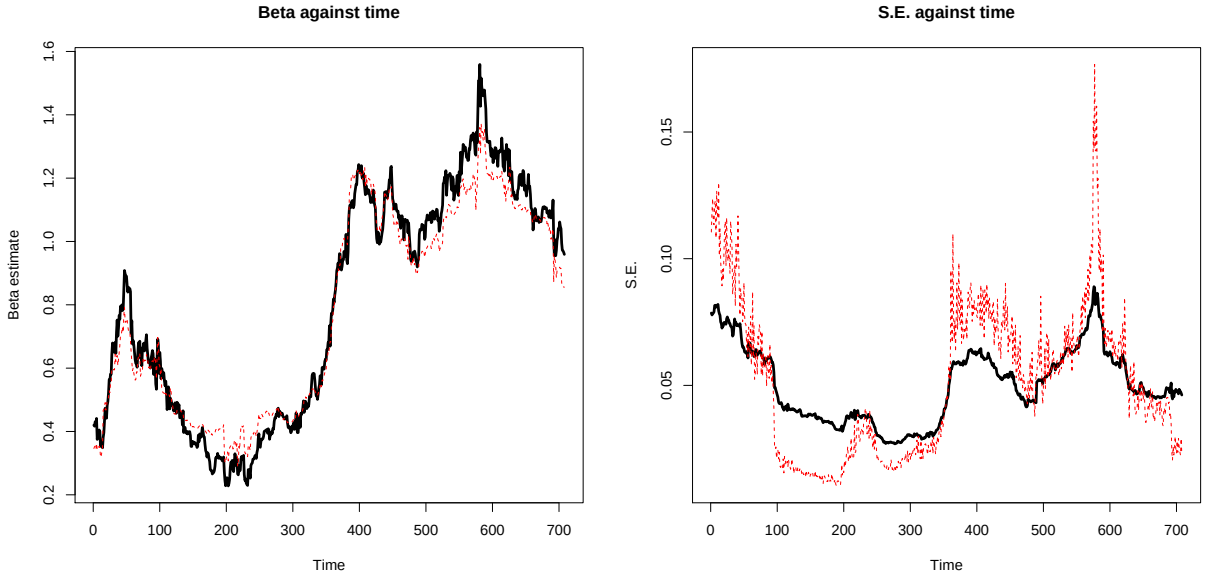

Figure 9: LHS shows the pair ($\hat{\beta}$,$\hat{\beta}_{LS}$) through time for ABT, Abbott Laboratories. Black line is $\hat{\beta}$, red dotted line is $\hat{\beta}_{LS}$. The RHS shows the pair $(SE(\hat{\beta}), SE(\hat{\beta}_{LS}))$ through time. Black line is $SE(\hat{\beta})$, red dotted line is $SE(\hat{\beta}_{LS})$.

around dramatically, easily upended by a single pair of data. More solid inference can be carried out using $\hat{\beta}$.

6.6 Rolling low and high beta portfolios

Using hypothesis tests based on the rolling estimates ($\hat{\beta}$,$\hat{\beta}_{LS}$), and corresponding standard errors $(SE(\hat{\beta}), SE(\hat{\beta}_{LS}))$, I built each week an equal weighted high beta and low beta portfolio. The stocks are selected using a hypothesis test, which means the number of stocks in portfolio is random — an alternative would be to put a fixed number of stocks in the portfolio, selecting the stocks with the lowest p-values. These portfolios are then used through the next week's data to form a return over a week. This procedure is run through the entire sample period. Table 4 reports out of sample summary statistics of these two portfolios.

The results for the portfolio selected using test statistics based on $\hat{\beta}$ and $\hat{\beta}_{LS}$ are similar.

The high beta portfolios produce what is expected: a higher average return and a substantially higher standard deviation, compared to the index. Their Sharpe ratios are not very different than the index, while their own betas are high, around 1.75, while the corresponding alphas are statistically close to 0.

The statistic "Share" is the time series average of the proportion of the cross-section which is in the portfolio. Hence the low beta portfolio contains around 13% of the stocks, the high beta portfolios contain around 7%. The "$|\Delta Share|$" is the time series average of the difference in holdings, so if the universe of stocks is 1,000 and the "$|\Delta Share| = 0.001$" then on average 1 stock moves into or out of the portfolio each week. In our case the universe of stocks is 358, so on average 3 stocks come into or out of the portfolios each week.

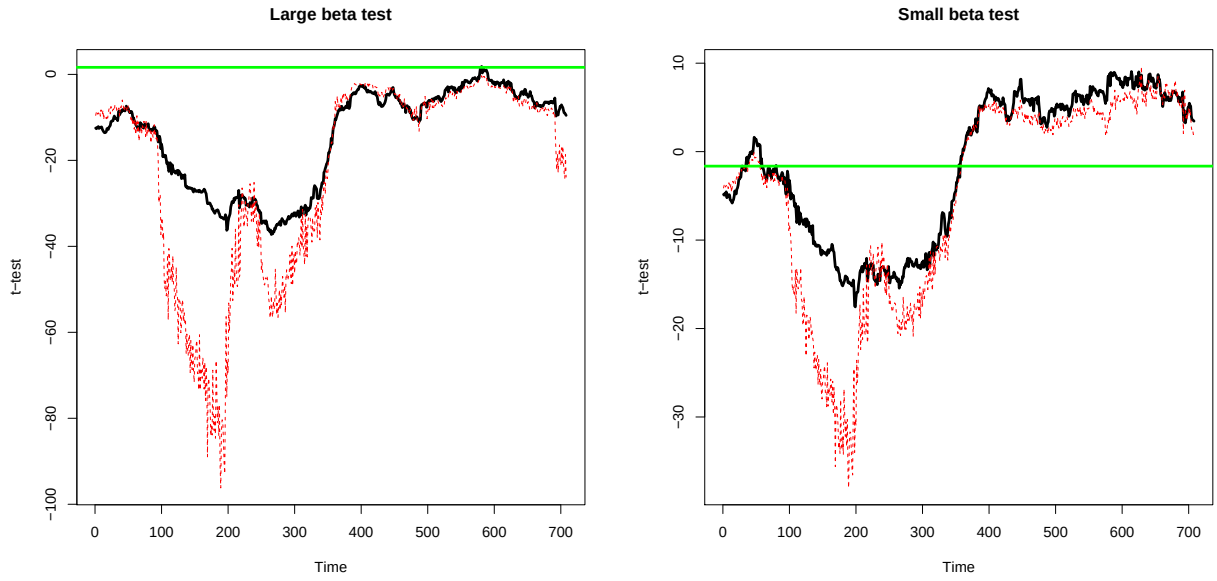

Figure 10: LHS shows $(\widehat{T}_{\widehat{\beta}}, \widehat{T}_{LS})$ of large beta stocks for ABT, Abbott Laboratories. Black line is $\widehat{T}_{\widehat{\beta}}$, red dotted line is $\widehat{T}_{LS}$. RHS shows $(\widehat{T}_{\widehat{\beta}}, \widehat{T}_{LS})$ of small beta stocks. The green lines are the critical values for the 5% nominal level tests.

Portfolio type	E	sd	Sharpe	alpha	beta	Share	$\lvert\Delta Share\rvert$
Index	0.194	2.49	0.077	0	1	0	0
Low beta $\widehat{\beta}$	0.213	1.96	0.109	0.093 (0.041)	0.625	0.136	0.008
Low beta $\widehat{\beta}_{LS}$	0.222	1.88	0.118	0.111 (0.042)	0.573	0.137	0.008
High beta $\widehat{\beta}$	0.361	5.05	0.071	0.018	1.77	0.076	0.007
High beta $\widehat{\beta}_{LS}$	0.359	5.09	0.070	0.018	1.75	0.076	0.006

Table 4: Week by week out of sample portfolio returns. Low and high beta portfolios are estimated by hypothesis testing. Figures in brackets are standard errors.

The low beta portfolios are more interesting economically: their low beta should deliver a portfolio with low average returns and low standard deviations, but actually the average returns are higher than for the index, and this pushes the Sharpe ratio up. When these portfolios are regressed against the index, the betas are around two thirds, while they have positive alpha (the figure in brackets are standard errors). This is certainly not an unexpected result from the applied finance literature. The low beta premium is reported extensively in the literature, see Baker et al. (2011) for a review.

7 Conclusions

I have advocated estimating linear in parameters predictive regressions using $\hat{\beta}$ rather than the celebrated least squares estimators $\hat{\beta}_{LS}$. Why? $\hat{\beta}$ has relatively simple standard errors which are robust to thick tailed predictors. This robustness leads, in practice, to nominal tests which

are close to being valid (i.e. correct size), as well as standard errors which a reasonable smooth through time for rolling estimators.

There are downsides with $\hat{\beta}$. $\hat{\beta}_{LS}$ gains precise from being super sensitive to unusual predictors. $\hat{\beta}$ downweights these low probability but influential predictors. Sometimes it is good to take full advantage of every scrap of available information. In that case it makes sense to use $\hat{\beta}_{LS}$. But replication and testing in that environment will be fragile, sensitive to one or two datapoints. $\hat{\beta}$ is a more conservative estimator, it even works with data as thick tailed as nearly Cauchy. Empirically in finance, the average standard deviation for $\hat{\beta}$ is close to $\hat{\beta}_{LS}$, so the practical loss of precision seems modest.

Finally, it has not escape my notice that the usual estimator of quantile regression has the same vulnerability to thick tailed predictors as least squares and some of the ideas expressed here can be used to tackle that problem too. I will address this in a further paper.

8 Acknowledgements

Thanks to Isaiah Andrews, Joseph Blitzstein, John Campbell, Roger Koenker, Nour Meddahi, Andrew Patton, Ashesh Rambachan, Julia Shephard, Roy Welsch and particularly Jihyun Kim. The code which produces all the simulation results is in `sign1.r`. The code which produces all the empirical results (including downloading the data) is in `RegFinance1.r` and `Recursive1.r`. An earlier version of this paper had corresponding material on quantile regression, but that is now being written up separately.

9 Appendix

9.1 Proof of Theorem 3

Now

$$\begin{aligned} S_{\mathbf{G},\mathbf{X}} &= \frac{1}{n}\sum_{j=1}^{n} \left\|\mathbf{X}_j\right\|_2^{-1} \begin{pmatrix} 1 & \left(\mathbf{Z}_j - \overline{\mathbf{Z}}\right)^{\intercal} \\ \mathbf{Z}_j - \overline{\mathbf{Z}} & \left(\mathbf{Z}_j - \overline{\mathbf{Z}}\right)\left(\mathbf{Z}_j - \overline{\mathbf{Z}}\right)^{\intercal} \end{pmatrix} = \begin{pmatrix} S_1 & S_{\mathbf{Z}-\overline{\mathbf{Z}}}^{\intercal} \\ S_{\mathbf{Z}-\overline{\mathbf{Z}}} & S_{\mathbf{Z}-\overline{\mathbf{Z}},\mathbf{Z}-\overline{\mathbf{Z}}} \end{pmatrix} \\ &= S_1 \begin{pmatrix} 1 & \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}}}^{\intercal} \\ \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}}} & \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}},\mathbf{Z}-\overline{\mathbf{Z}}} \end{pmatrix} = S_1 \begin{pmatrix} 1 & \left(\widetilde{\mathbf{Z}} - \overline{\mathbf{Z}}\right)^{\intercal} \\ \widetilde{\mathbf{Z}} - \overline{\mathbf{Z}} & \widetilde{S}_{\mathbf{ZZ}} - \widetilde{\mathbf{Z}}\overline{\mathbf{Z}}^{\intercal} - \overline{\mathbf{Z}}\widetilde{\mathbf{Z}}^{\intercal} + \overline{\mathbf{Z}}\overline{\mathbf{Z}}^{\intercal} \end{pmatrix}, \end{aligned}$$

and

$$S_{\mathbf{G},Y} = \frac{1}{n}\sum_{j=1}^{n} \left\|\mathbf{X}_j\right\|_2^{-1} \begin{pmatrix} Y_j \\ \left(\mathbf{Z}_j - \overline{\mathbf{Z}}\right) Y_j \end{pmatrix} = S_1 \begin{pmatrix} \widetilde{S}_Y \\ \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}},Y} \end{pmatrix} = S_1 \begin{pmatrix} \widetilde{Y} \\ \overline{S}_{\mathbf{Z}Y} - \widetilde{Y}\overline{\mathbf{Z}} \end{pmatrix}$$

Now

$$\begin{pmatrix} 1 & \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}}}^{\intercal} \\ \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}}} & \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}},\mathbf{Z}-\overline{\mathbf{Z}}} \end{pmatrix} \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_{1:p} \end{pmatrix} = \begin{pmatrix} \widetilde{S}_Y \\ \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}},Y} \end{pmatrix}$$

so that

$$\widehat{\beta}_0 + \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}}}^{\intercal}\widehat{\beta}_{1:p} = \widetilde{S}_Y$$

$$\widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}}}\widehat{\beta}_0 + \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}},\mathbf{Z}-\overline{\mathbf{Z}}}\widehat{\beta}_{1:p} = \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}},Y}.$$

Rearranging gives the result for $\widehat{\beta}_0$. Note that

$$\begin{aligned}\widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}},\mathbf{Z}-\overline{\mathbf{Z}}} - \widetilde{S}_{\mathbf{Z}-\overline{\mathbf{Z}}}\widetilde{S}^{\mathrm{T}}_{\mathbf{Z}-\overline{\mathbf{Z}}} &= \widetilde{S}_{\mathbf{ZZ}} - \widetilde{\mathbf{Z}}\overline{\mathbf{Z}}^{\intercal} - \overline{\mathbf{Z}}\widetilde{\mathbf{Z}}^{\mathrm{T}} + \overline{\mathbf{Z}}\overline{\mathbf{Z}}^{\intercal} - \left(\widetilde{\mathbf{Z}} - \overline{\mathbf{Z}}\right)\left(\widetilde{\mathbf{Z}} - \overline{\mathbf{Z}}\right)^{\mathrm{T}} \\ &= \widetilde{S}_{\mathbf{Z}-\widetilde{\mathbf{Z}},\mathbf{Z}-\widetilde{\mathbf{Z}}}.\end{aligned}$$

Then

$$\begin{aligned}\overline{S}_{\mathbf{Z}-\overline{\mathbf{Z}},\mathbf{Z}-\overline{\mathbf{Z}}}\widehat{\beta}_{1:p} &= \overline{S}_{\mathbf{Z}-\overline{\mathbf{Z}},Y} - \overline{S}_{\mathbf{Z}-\overline{\mathbf{Z}}}\widehat{\alpha} = \overline{S}_{\mathbf{Z}-\overline{\mathbf{Z}},Y} - \overline{S}_{\mathbf{Z}-\overline{\mathbf{Z}}}\overline{S}_Y + \overline{S}_{\mathbf{Z}-\overline{\mathbf{Z}}}\overline{S}^{\intercal}_{\mathbf{Z}-\overline{\mathbf{Z}}}\widehat{\gamma} \\ \left(\overline{S}_{\mathbf{Z}-\overline{\mathbf{Z}},\mathbf{Z}-\overline{\mathbf{Z}}} - \overline{S}_{\mathbf{Z}-\overline{\mathbf{Z}}}\overline{S}^{\intercal}_{\mathbf{Z}-\overline{\mathbf{Z}}}\right)\widehat{\beta}_{1:p} &= \overline{S}_{\mathbf{Z}-\overline{\mathbf{Z}},Y} - \overline{S}_{\mathbf{Z}-\overline{\mathbf{Z}}}\overline{S}_Y \\ \overline{S}_{\mathbf{Z}-\widetilde{\mathbf{Z}},\mathbf{Z}-\widetilde{\mathbf{Z}}}\widehat{\beta}_{1:p} &= \overline{S}_{\mathbf{Z}-\widetilde{\mathbf{Z}},Y-\widetilde{Y}},\end{aligned}$$

then using a matrix inverse gives the result.

9.2 Proof of Theorem 4

B1 is needed for $\widehat{\beta}$ to exist. Define the predictive regression errors $U_j = Y_j - \mathbf{X}_j^{\intercal}\beta$, for $j = 1, ..., n$. Then

$$\widehat{\beta} = S^{-1}_{\mathbf{G},\mathbf{X}}S_{\mathbf{G},Y} = \beta + S^{-1}_{\mathbf{G},\mathbf{X}}\frac{1}{n}\sum_{j=1}^{n}\mathbf{G}_j U_j.$$

Then under A2 and B2,

$$\mathrm{E}[U_j|(\mathbf{Z} = \mathbf{z})] = 0,$$

while under B2 and B3

$$\mathrm{Var}(U_j|(\mathbf{Z} = \mathbf{z})) = \sigma_j^2 < \infty.$$

Then the conditional unbiasedness and conditional variance follow.

What remains is the Berry-Esseen type result. The approach is to use the Bentkus (2005) version. It is stated here for convenience.

Theorem 6 *(Bentkus (2005)) Suppose W_i are zero mean, independent in R^d, then*

$$\sup_{A\in\mathcal{C}^d}\left|\Pr(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}V_j \in A) - \Pr(N(0, C^*) \in A)\right| \leq c\frac{d^{1/4}}{n^{1/2}}\left(\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}[\left\|(C^*)^{-1/2}V_j\right\|_2^3]\right),$$

where

$$C^* = n^{-1}\sum_{j=1}^{n}\mathrm{Var}(V_j),$$

where $\mathcal{C}^d$ denotes the set of all convex subsets of R^{p+1}

Now

$$\sqrt{n}(\widehat{\beta}-\beta)=\sum_{j=1}^{n}V_j,\quad \text{where}\quad V_j=\frac{1}{\sqrt{n}}S_{\mathbf{G},\mathbf{X}}^{-1}\mathbf{G}_jU_j.$$

The V_j are conditionally independent with variances

$$C_j=\mathrm{Var}(V_j|(\mathbf{Z}=\mathbf{z}))=\frac{1}{n}\sigma_j^2S_{\mathbf{G},\mathbf{X}}^{-1}\mathbf{G}_j\mathbf{G}_j^TS_{\mathbf{G},\mathbf{X}}^{-1},$$

and the corresponding sums

$$V^*=\sum_{j=1}^{n}V_j,\quad C^*=\sum_{j=1}^{n}C_j=S_{\mathbf{G},\mathbf{X}}^{-1}S_{\sigma^2\mathbf{G},\mathbf{G}}S_{\mathbf{G},\mathbf{X}}^{-1}.$$

C^* is invertible using B1 and B4. Then

$$(C^*)^{-1}=S_{\mathbf{G},\mathbf{X}}S_{\sigma^2\mathbf{G},\mathbf{G}}^{-1}S_{\mathbf{G},\mathbf{X}},$$

which is positive definite, so

$$(C^*)^{-1/2}=S_{\sigma^2\mathbf{G},\mathbf{G}}^{-1/2}S_{\mathbf{G},\mathbf{X}},\quad \text{that is}\quad \left\{(C^*)^{-1/2}\right\}^{\mathtt{T}}(C^*)^{-1/2}=(C^*)^{-1}$$

Then define the corresponding skewness terms

$$\varsigma_j=\mathrm{E}\left\|(C^*)^{-1/2}V_j|(\mathbf{Z}=\mathbf{z})\right\|_2^3,\quad \varsigma^*=\sum_{j=1}^{n}\varsigma_j,$$

which exists using B5.

Then

$$\begin{aligned}(C^*)^{-1/2}V_j &= S_{\sigma^2\mathbf{G},\mathbf{G}}^{-1/2}S_{\mathbf{G},\mathbf{X}}V_j=\frac{1}{\sqrt{n}}S_{\sigma^2\mathbf{G},\mathbf{G}}^{-1/2}S_{\mathbf{G},\mathbf{X}}S_{\mathbf{G},\mathbf{X}}^{-1}\mathbf{G}_jU_j\\ &= \frac{1}{\sqrt{n}}S_{\sigma^2\mathbf{G},\mathbf{G}}^{-1/2}\mathbf{G}_jU_j,\end{aligned}$$

so

$$\varsigma_j=\mathrm{E}[|U_j/\sigma_j|^3|(\mathbf{Z}=\mathbf{z})]\left\|S_{\sigma^2\mathbf{G},\mathbf{G}}^{-1/2}\mathbf{G}_j\sigma_j\right\|_2^3.$$

Then the result follows from Bentkus (2005).

9.3 Proof of Theorem 5

Stack the moment conditions:

$$g(Y_1,\mathbf{Z}_1;\theta)=\begin{pmatrix}g_1(\mathbf{Z}_1;\psi)\\ g_2(Y_1,\mathbf{Z}_1;\psi,\beta)\end{pmatrix},\quad g_1(\mathbf{Z}_1;\psi)=\mathbf{Z}_1-\psi,\quad g_2(Y_1,\mathbf{Z}_1;\psi,\beta)=\mathbf{G}_1(\psi)\left\{Y_1-\mathbf{X}_1(\psi)^{\mathtt{T}}\beta\right\}$$

where $\theta=(\psi^{\mathtt{T}},\beta^{\mathtt{T}})^{\mathtt{T}}=(\psi^{\mathtt{T}},\alpha,\gamma^{\mathtt{T}})^{\mathtt{T}}$. Crucially g_1 does not depend upon β. Then

$$\mathrm{E}[g(Y_1,\mathbf{Z}_1;\theta_0)]=0,$$

while

$$\mathrm{Var}[g(Y_1, \mathbf{Z}_1; \theta_0)] = \begin{pmatrix} \mathrm{Var}(\mathbf{Z}_1) & 0 \\ 0 & \mathrm{Var}(\mathbf{G}_1 U_1) \end{pmatrix},$$

where here I write $\mathbf{G}_1 = \mathbf{G}_1(\psi_0)$ and $\mathbf{X}_1(\psi) = \mathbf{X}_1(\psi_0)$. Of course

$$\mathrm{Var}(\mathbf{G}_1 U_1) = \mathrm{E}(\sigma^2 \mathbf{G}_1 \mathbf{G}_1^{\mathsf{T}}).$$

The crucial matrix zeros follow by Adam's law applied to

$$\mathrm{E}[(\mathbf{Z}_1 - \psi)\mathbf{G}_1 U_1] = \mathrm{E}[(\mathbf{Z}_1 - \psi)\mathbf{G}_1 \mathrm{E}[U_1 | Z_1]] = 0.$$

Further,

$$\left.\frac{\partial \mathrm{E}[g(Y_1, \mathbf{Z}_1; \theta)]}{\partial \theta^T}\right|_{\theta=\theta_0} = \begin{pmatrix} -I_p & 0 \\ V_0 & -\mathrm{E}[\mathbf{G}_1 \mathbf{X}_1^{\mathsf{T}}] \end{pmatrix}.$$

The term V_0 is not of central importance, but takes the form

$$\begin{aligned} V_0 &= -\mathrm{E}[\mathbf{G}_1(\psi_0)\beta^{\mathsf{T}} \frac{\partial \mathbf{X_1}(\psi_{\mathbf{0}})}{\partial \psi^{\mathsf{T}}}], \quad \frac{\partial \mathbf{X}_1(\psi_0)}{\partial \psi^{\mathsf{T}}} = (0_p, -I_p) \\ &= -\mathrm{E}[\mathbf{G}_1(\psi)]\beta^{\mathsf{T}} (0_p, -I_p) = \mathrm{E}[\mathbf{G}_1(\psi)]\gamma^{\mathsf{T}}. \end{aligned}$$

Finally, notice A3 means that $\mathrm{E}[\mathbf{G}_1 \mathbf{X}_1^{\mathsf{T}}]$ is symmetric and invertible. The result then follows conventionally.

QED.

References

Andrews, I., J. Stock, and L. Sun (2019). Weak instruments in IV regression: Theory and practice. *Annual Review of Economics 11*, 727–753.

Baker, M., B. Bradley, and J. Wurgler (2011). Benchmarks as limits to arbitrage: Understanding the low-volatility anomaly. *Financial Analysts Journal 67*, 1–15.

Balkema, G. and P. Embrechts (2018). Linear regression for heavy tails. *Risks 6*, 1–70.

Baringhouse, L. and N. Henze (2017). Cramér–von Mises distance: probabilistic interpretation, confidence intervals, and neighbourhood-of-model validation. *Journal of Nonparametric Statistics 29*, 167–188.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2011). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics 162*, 149–169.

Barndorff-Nielsen, O. E. and N. Shephard (2004). Econometric analysis of realised covariation: high frequency covariance, regression and correlation in financial economics. *Econometrica 72*, 885–925.

Bentkus, V. (2005). A Lyapunov-type bound in R^d. *Theory of Probability and its Applications 49*, 311–323.

Black, F. (1972). Capital market equilibrium with restricted borrowing. *Journal of Business 4*, 444–455.

Blattberg, R. and T. J. Sargent (1971). Regression with non-Gaussian disturbances: Some sampling results. *Econometrica 39*, 501–510.

Bollerslev, T., A. Patton, J. Li, and R. Quaedvlieg (2020). Realized semicovariances. *Econometrica*. Forthcoming.

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

Cauchy, A. (1836). On a new formula for solving the problem of interpolation in a manner applicable to physical investigations. *Philosophical Magazine 8*, 459–468.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance 1*, 223–236.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 59–82. Berkeley: University of California.

Engle, R. F. (2016). Dynamic conditional beta. *Journal of Financial Econometrics 14*, 643–667.

Farebrother, R. W. (1999). *Fitting Linear Relationships: A history of the calculus of observations 1750–1900*. New York: Springer.

Gorji, F. and M. Aminghafari (2019). Robust nonparametric regression for heavy-tailed data. *Journal of Agricultural, Biological, and Environmental Statistics*. Forthcoming.

Hall, P. G. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer.

Hallin, M., Y. Swan, T. Verdebout, and D. Veredas (2010). One-step R-estimation in linear models with stable errors. *Journal of Econometrics 172*, 195–204.

Hampel, F., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (2005). *Robust Statistics: The Approach based on Influence Functions*. New York: Wiley.

Hausman, J. and C. Palmer (2012). Heteroskedasticity-robust inference in finite samples. *Economics Letters 116*, 232–235.

Heyde, C. C. and E. Seneta (1977). *I.J. Bienayme. Statistical Theory Anticipated*. Springer. Studies in the History of Mathematics and Physical Sciences, volume 3.

Hill, J. B. and E. Renault (2010). Generalized method of moments with tail trimming. Unpublished paper: Department of Economics, UNC.

Hollstein, F., M. Prokopczuk, and C. W. Simen (2019). Estimating beta: Forecast adjustments and the impact of stock characteristics for a broad cross-section. *Journal of Financial Markets 44*, 91–118.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 221–233. University of California Press.

Ibragimov, R., J. Kim, and A. Shrobotov (2020). New robust inference for predictive regressions. Unpublished paper: Imperial College, London.

Karolyi, G. A. (1992). Predicting risk: some new generalizations. *Management Science 38*, 57–74.

Kim, J. and N. Meddahi (2020). Volatility regressions with fat tails. *Journal of Econometrics*. Forthcoming.

King, G. and M. E. Roberts (2015). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis 23*, 159–179.

Krasker, W. S. and R. E. Welsch (1982). Efficient bounded-influence regression estimation. *Journal of the American Statistical Association 77*, 595–604.

Kurz-Kim, J. and M. Loretan (2014). A note on the coefficient of determination in regression models with infinite-variance variables. *Journal of Econometrics 181*, 15–24.

Linnik, Y. (1961). *Method of Least Squares and Principal Theory of Observations*. New York: Pergamon Press.

MacKinnon, J. G. and H. White (1985). Some heteroskedastic-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29*, 305–325.

MacKinnon, R. (2012). Thirty years of heteroskedasticity-robust inference. In X. Chen and N. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 437–461. Springer.

Mallows, C. L. (1975a). Influence functions. Presented at the NBER conference on Robust Regression, Cambridge, MA.

Mallows, C. L. (1975b). On some topics in robustness. Unpublished paper: Bell Telephone Laboratories, Murray Hill, New Jersey.

Mikosch, T. and C. G. de Vries (2013). Heavy tails of OLS. *Journal of Econometrics 172*, 205–21.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden (Eds.), *The Handbook of Econometrics, Volume 4*, pp. 2111–2245. North-Holland.

Nolan, J. P. and D. Ojeda-Revah (2013). Linear and nonlinear regression with stable errors. *Journal of Econometrics 172*, 186–194.

Phillips, P. C. B., J. Y. Park, and Y. Chang (2004). Nonlinear instrumental variable estimation of an autoregression. *Journal of Econometrics 118*, 219–246.

Samorodnitsky, G., S. T. Rachev, J.-R. Kurz-Kim, and S. V. Stoyanov (2007). Asymptotic distribution of unbiased linear estimators in the presence of heavy-tailed stochastic regressors and residuals. *Probability and Mathematical Statistics 27*, 275–302.

Seal, H. L. (1967). Studies in the history of probability and statistics. XV: The historical development of the Gauss linear model. *Biometrika 67*, 1–24.

So, B. S. and D. W. Shin (1999). Cauchy estimators for autoregressive processes with applications to unit root tests and confidence intervals. *Econometric Theory 15*, 165–176.

Stambaugh, R. F. (1999). Predictive regression. *Review of Financial Economics 54*, 375–421.

Sun, Q., W. Zhou, and J. Fan (2020). Adaptive Huber regression. *Journal of the American Statistical Association 115*, 254–265.

Vasicek, O. A. (1973). A note on using cross-sectional information in Bayesian estimation of security betas. *Journal of Finance 28*, 1233–1239.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica 48*, 817–838.