VERSITA

Folia Oeconomica Stetinensia
DOI: 10.2478/v10031-012-0032-7

Wydział Nauk Ekonomicznych i Zarządzania
Uniwersytetu Szczecińskiego

SELECTED ROBUST METHODS FOR CAMP MODEL ESTIMATION

Grażyna Trzpiot, Prof.

University of Economics in Katowice,
Faculty of Informatics and Communication
Department of Demography and Economic Statistics
Bogucicka 14, 40-226 Katowice, Poland
e-mail: grazyna.trzpiot@ue.katowice.pl

Received 15 November 2012, Accepted 23 April 2013

Abstract

This paper presents evidence that Ordinary Least Squares estimators of beta coefficients of major firms and portfolios are highly sensitive to observations of extremes in market index returns. This sensitivity is rooted in the inconsistency of the quadratic loss function in financial theory. By introducing considerations of risk aversion into the estimation procedure using alternative estimators measures of variability we can overcome this lack of robustness and improve the reliability of the results.

Keywords: OLS estimators, systematic risk, LTS estimators, quantile regression estimators.

JEL classification: G12, C14.

Introduction

The valuation of risky assets is one of the major research tasks in financial economics that has led to the development of several Capital Asset Pricing Models, the most popular of which is the Sharpe-Lintner-Black mean-variance CAPM. In this model, the typical measure of asset riskiness is the beta, i.e., the covariance between the asset return and the market portfolio return. The basic tenet of CAPM lies in the separation of estimating beta risk from its pricing. Indeed CAPM assumes that one can define and measure systematic risk irrespective of risk aversion, which affects only the equilibrium pricing of individual assets. As is well known, this separation is valid only under the restrictive assumption of two-factor separating distributions or alternatively, if the utility function is quadratic.

Empirical asset-pricing models attract massive attention in finance, their goal being to assert or refute whether CAPM holds true. The traditional technique used to estimate the risk-expected return relation consists of two stages. In the first pass, betas are estimated from a time-series. In the second pass, the relationship between mean returns and betas is tested across firms or portfolios. Since its inception in finance, beta has been used mainly for two purposes. The first involves the ranking of assets and portfolios with respect to systematic risk by practitioners. The second deals with testing CAPM and mean-variance efficiency. In this paper we question whether the standard procedure for estimating systematic risk is compatible with financial theory and show how the regression technique used to estimate systematic risk is not robust with respect to wide market fluctuations. The sensitivity of beta to the presence of extreme observations can give rise to data mining and lead the way to peculiar relationships.

The goal of this paper is to present selected robust methods for the CAPM model estimation. The proposed approach has been applied to a selected part of Polish financial markets.

The paper is organized as follows. Section 1 presents the *OLS* estimator for beta as a weighted average of the change in asset return conditional on the change in market returns. The weights used in averaging depend solely on the distribution of market returns. As the weights are sensitive to extreme market fluctuations, the *OLS* estimation procedure attaches greater weights to extreme market changes, a characteristic that may contradict financial theory. In Section 2, we show chosen robust linear regression model as LTS and QR model. In Section 3 we discus how to deal with outliers in multivariate regression analysis, and how it can influence the results of analysis. In Section 4, we offer selected robust methods for market model estimation, alternative estimators for describing the riskiness of an asset such as LTS and QR model and investigate their properties. These estimators attach lower weights than the *OLS*

estimators to upward market movements, thus making the estimator both more appropriate from the theoretical point of view, and at the same time more robust than the *OLS* estimator. Closing section concludes the paper.

1. Estimation of financial market equilibrium model

Estimator of model slope *beta* determined with the use of LS is a weighted mean of slopes obtained from two nearest-neighbour observations across security characteristic curve. This renders impossible verifying what weights are assigned to extreme values of rate of return in the sample[1]. Let us consider a market model where the rate of return on investment is random and continuous described by density function $f(R_k, M)$, where R_k is return on equity k and M is market portfolio. Let f_M, F_M, μ_M, and σ^2_M be margin density, margin distribution, expected value and variance M respectively. We assume that the first and second moment exist and we define $R_k(m) = E(R_k| M = m)$ as conditional market rate of return k assuming that market rate of return $M = m$. The value $R_k(m)$ determines security characteristic curve[2]. With the aim of estimating *beta* of a stock, we determine the following relationship referred to as CAMP model:

$$R_k = \alpha_k + \beta_k M + \varepsilon_k \tag{1}$$

Additional assumption being that random component ε_k is independent, identically distributed, of expected value equal zero and constant variance, the LS estimator is given by[3]:

$$\beta_{LS} = \frac{\operatorname{cov}(R_k, M)}{\operatorname{cov}(M, M)} \tag{2}$$

When estimating linear model parameters, as well as in risk estimation defined as *beta* it is important to correctly assume the error distribution. Should error distribution have the Gaussian distribution, then the LS estimator of the model parameter has minimum variance in unbiased estimator class[4]. By using Jensen's inequality and assuming normal distribution, the optimisation procedure for the LS estimators can be employed for any convex loss function.

Should error distribution could not be approximated by the Gaussian distribution using LS then we get the best unbiased estimator of linear model only once we concentrate on parameters being linear function of the dependent variable. In many cases that set could be unnecessarily restrictive.

By deploying statistical modelling, fat-tailed distribution could be modelled as a combination of normal distributions. For instance analysed data can be generate from standard

normal distribution, but could be disturbed by observations from normal distributions with higher variance. Such distribution would have fatter tails than normal distribution[5].

Financial literature shows early research provides observation that daily rates of return have fat-tailed distribution. Fama[6] applied stable Pareto distribution to daily observations and concluded that characteristic exponent of distribution was less than 2. Another paper discussed the student's t distribution[7]. Kon formulated rate of return for Dow Jones Industrial Average using two up till four normal distributions[8]. Summarising results of those empirical research it is concluded that residual distribution does not resemble normal distribution and is fat-tailed.

Roll put forward an economic model which used rates of return with mixture distribution[9]. In essence, he assumed rates of return are intermittent with extreme values related to latest news, yet increase kurtosis of rate of return distribution.

Robust statistical methods present different approach to LS, however, they are slow to penetrate the world of finance. Determined estimators allocated less weight to outliers, for instance by minimising the sum of absolute deviations (Minimum Absolute Deviation MAD method) instead of using sum of squared deviations. Sharpe, Cornell and Dietrich employed MAD at *beta* risk estimation[10]. They concentrated their effort on rates of return for biggest companies and investment funds. Their findings showed that differences between the two methods (LS and MAD) are inconsiderable and do not prove any particular method to be ahead of the other.

2. Robust linear regression model

Linear regression[11] was first defined in the 80's of last century. The very first most renowned regression was given by:

$$\min_{b} \underset{i}{median} \left| y_i - x_i b \right|^2 \tag{3}$$

and is referred to as least median of squares – LMS.

Justification for residual squares is an observation where n is even, then median centre is selected. That is a very robust regression which does not require parameter of scale estimation. Since it covers $1/\sqrt[3]{n}$ of data at most, it is very inefficient.

Ruppert and Carroll suggest regression of the *trimmed least squares* – LTS[12].

$$\min_{b} \sum_{i} \left| y_i - x_i b \right|^2_{(i)} \tag{4}$$

This method is far more efficient, but separates only extreme observations[13]. Recommended sum of residual squares should not exceed $q = [(n + p + 1)/2]$.

That approach was then replaced by S-estimators, for which regression equation coefficients are solved for solution to the problem

$$\sum_{i=1}^{n} \chi\left(\frac{y_i - x_i b}{c_o s}\right) = (n-p)\beta \tag{5}$$

with least s scale parameter. In the last equation the χ function is usually assumed as integrable Tukey's biweight function.

$$\chi(u) = \begin{cases} u^6 - 3u^4 + 3u^2, & |u| \le 1 \\ 1, & |u| \ge 1 \end{cases} \tag{6}$$

Values $c_0 = 1.548$ and $\beta = 0.5$ are selected for goodness of fit, should error distribution be normal distribution. That yields efficiency of 28.7% for normal distribution, which is low but still better that LS and LTS.

In least square method estimators are solved for through minimising sum of residual squares. Below-proposed estimator for minimisation uses the following criterion:

$$\sum_{t=1}^{T} \rho_\theta(u_t) \tag{7}$$

for $\rho_\theta(u_t) = \theta|u_t|$, if $u_t \ge 0$ or $\rho_\theta(u_t) = (1-\theta)|u_t|$ if $u_t < 0$,

where $0 < \theta < 1$, $u_t = r_t - \alpha - \beta r_{mt}$, $t = 1, ..., T$.

Since minimad (MAD) is the sum of absolute deviations of residuals, observations are considered differently to the sum of residual squares[14]. In general, high (low) value of "weight" θ yields high observation penalty with substantial negative (positive) residual. Each regression line fitted (corresponding to values different than θ) intersects at least two points from the pool of data, with highest T number θ of observations from sample beneath fitted line, and at least $(T - 2)\theta$ observations above that line[15]. Considering values θ from interval <0, 1> we get a set of regression quantile estimators $\hat{\beta}(\theta)$, resembling sampling quantile distribution for sampling quantile distribution[16].

That very specific effect or positive or negative outlier will determine quantile regression corresponding to extreme (either high or low) value of θ. One should remember, however, that no observations are removed during processing of statistics. Furthermore, volatility of rate of return determines changing of quantile regression for different θ values. From this perspective

using β as estimator corresponding to one θ value, with the MAD estimation method, could loss some useful information from sample[17]. Behaviour of estimators determined through MAD could be expanded by introducing an estimator based on bundle of quantile regression. Statistics related literature puts emphasis on producing robust estimator of mean population as linear combination of sample quantiles – trimmed means.

3. Outliers in multivariate regression analysis

In multivariate regression analysis outliers have no typical values of $Y(y_i)$ variable for corresponding variables $X(x_i)$ (vertical shift), and consequently produce high residuals (e_i). Outliers could also have inconsiderable residuals, but no typical values of explanatory variable. Those observations alter estimators thus results of multivariate regression analysis.

In simple regression (one explanatory variable) an observation with high y_i value for given x_i have high *discrepancy*, whereas observation with typical value of y_i variable for no typical value of x_i has high leverage and small residual (e_i). Observation with high leverage could have a small residual, but not necessarily. Observations with high leverage draw regression line towards y_i value. Consequently, influence of given observation of regression coefficient is expressed as function of discrepancy and leverage[18]. Diagnostics of observation's impact on multivariate regression analysis focus to outlier analysis or direct assessment of observation's influence on coefficients and fitting of determined regression model.

An observation is considered influential, should it considerably change model parameters due to inconsiderable change in its value or removal from sample. Residuals for typical observations are not high. Characteristic for outliers are high residuals i.e. difference between empirical value and theoretical value produced by estimated regression model[19].

An outlier is an observation considerably different to other. Normally it is caused by atypical factors. In the least square method, such single observation is capable of substantially changing estimated regression equation. In case of simple regression outlier could be detected by employing graphical analysis. Characteristics for outliers are high residuals. Hence it could become a whistle-blower detecting outliers, however, it shows certain shortcomings:

- residual are denominated quantities, whereas a good measure should be nominal – universal for all variables,
- no possibility of comparing residuals with independent template and thus difficulties with unambiguously ascertaining whether a residual is high or not.

Hence standardization of residuals is proposed. In literature concerning regression diagnostics, we encounter three methods for determining standardised residuals[20]:

1) $\tilde{e}_i = \dfrac{e_i}{s}$
   where $s^2 = e^T e \,/\, (n-k-1)$ is a classic estimator σ^2;
2) $e_i^* = \dfrac{ei}{s\sqrt{1-h_i}}$;
3) $e_{(i)}^* = \dfrac{ei}{s(i)\sqrt{1-h_i}}$
   where $s_{(i)}$ is estimation of standard deviation of random component σ after removal of i-th observation, h_i is element of diagonal projections matrix[21].

Expression from denominator of second standardised residual e_i^* is an estimator of standard deviation of normal residual e_i. Similarly interpreted is denominator of third residual $e_{(i)}^*$, whose premise involves removal of individual observations. Bear in mind, however, that standardised residual are not stochastically independent. Nevertheless residuals standardised through third method have student's t distribution with $n - 2 - k$ degrees of freedom. This is a key fact in multivariate regression analysis, since it allows statistical testing at predetermined significance level α. Hence they $e_{(i)}^*$ are referred to as *studentized residual* typically employed to detect outliers, which are the measure of observation's discrepancy.

Because of possible stochastic dependency between residuals, there are no reasons to discard i-th observation at significance level α, for $\left|e_i^*\right| > t_{n-2-k}\ (\alpha)$, it can be hold the boundary value of $\left|e_i^*\right| = 2$ or both approaches can be combined.

4. Selected robust methods for market model estimation

Empirical analysis of Sharpe model was attempted for companies listed under WIG20 stock market index. It was focused on biggest companies and observation period was from 13.07.2011 to 8.08.2012. Preliminary analysis of daily rates of returns on analysed assets showed presence of outliers (Figure 1) and extreme observations for all companies over the observation period. To further calibrate models of market rate of return selected were four companies (ticker) BOGDANKA (LWB), PGNIG (PGN), TAURONPE (TPE) and TPSA (TPS). They were chosen based on lowest value of coefficient of determination R^2 i.e. weakest match of linear models estimated by the least square method (Figure 2). For completeness of statistical analysis, Shapiro-Wilk test of normality of chosen variables were carried out, which confirmed they do not come from normal distribution (Table 1).

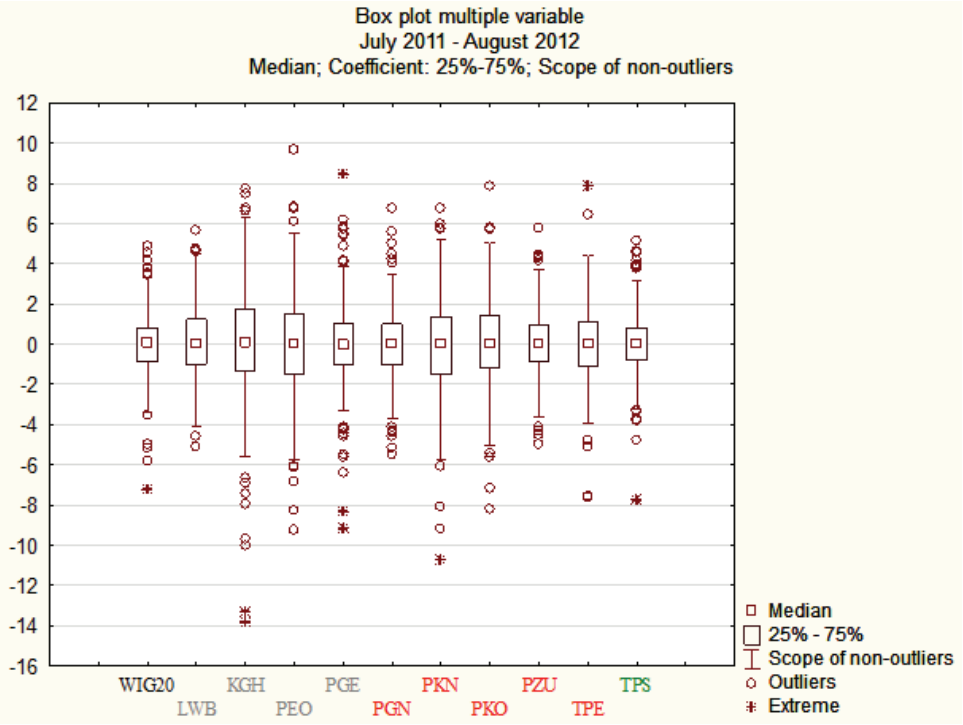

Fig. 1. Analysis of rate of return on stocks between 13.07.2011 and 8.08.2012

Source: own study.

Table 1. Results of Shapiro-Wilk test of normality

Asset	BOGDANKA (LWB)	PGNIG (PGN)	TAURONPE (TPE)	TPSA (TPS)
S-W test value	0.99016	0.98792	0.970330	0.96236
p-value	0.06066	0.02115	0.000002	0.00000

Source: own study.

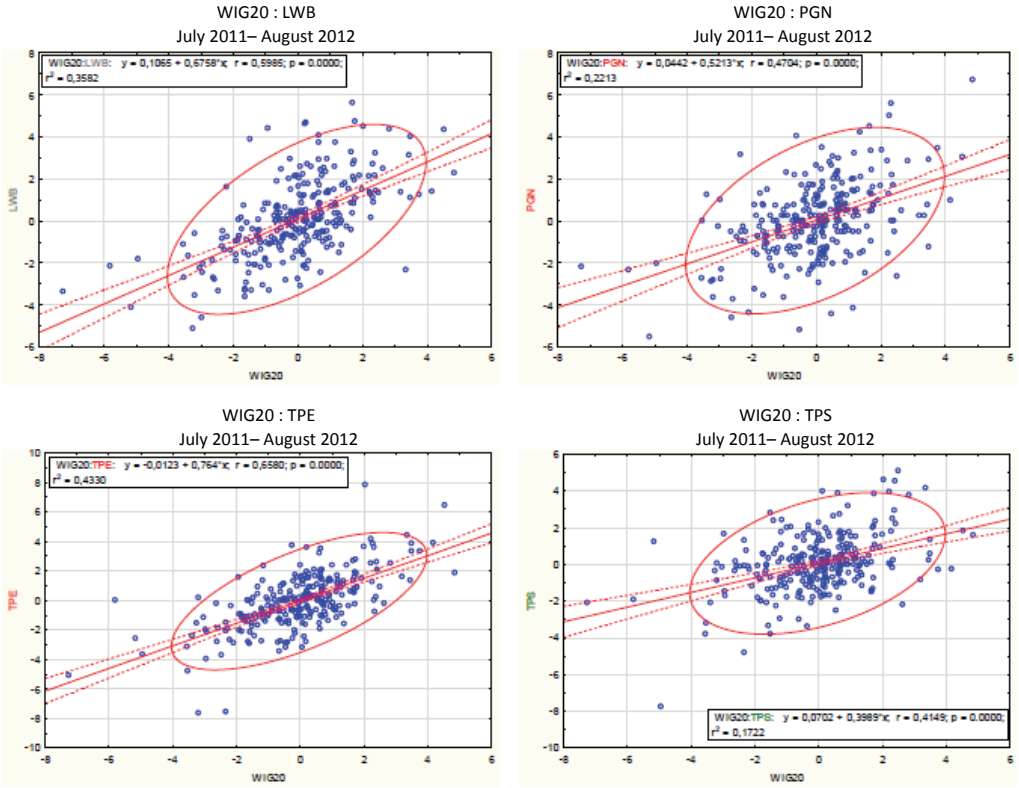

Fig. 2. Analysis of correlation and LS of companies (ticker) BOGDANKA (LWB), PGNIG (PGN), TAURONPE (TPE) and TPSA (TPS)

Source: own study.

Subsequently parameters for three chosen models were estimated. Classic LS model was compared with least trimmed squares methods LTS. Since the market model was by definition linear, and outliers analysis (Figure 3) did not confirm that assumption, additionally quantile regression was determined which corresponds the way of modelling which is different to asset pricing model. Tables 2–5 presents estimates of parameters for LS linear model, least trimmed squares LTS (residual analysis was used) and quantile regression model QR[22] for selected quantile level 0.01 ($VaR_{0,01}$) for the group of analysed companies. Diagnostics of influential observations executed for LTS model estimation provides information enabling reduction in number of observations and requires probing reliability of produced conclusions – key for further analysis of stock pricing model – which could be drawn based on fitted regression function. This also applies to influential observations distant from others, what gives basis to determine range of variable values the model can yield, for which conclusions should not be generalised.

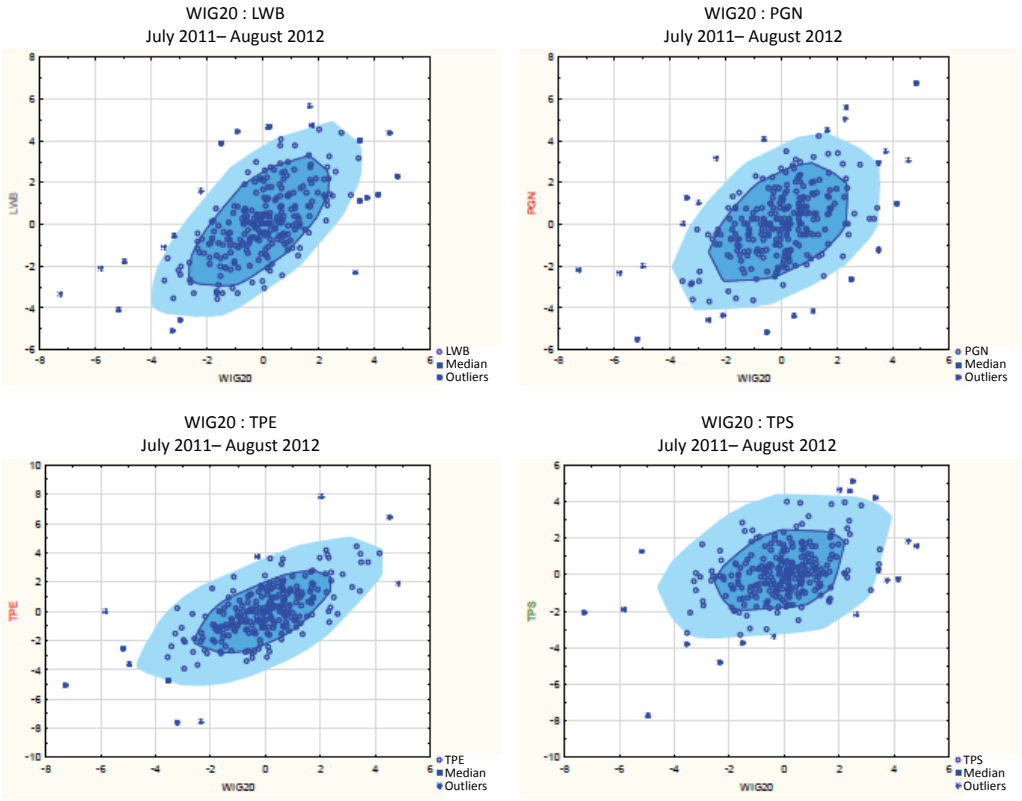

Fig. 3. LMS maps showing outliers for companies (ticker) BOGDANKA (LWB), PGNIG (PGN), TAURONPE (TPE) and TPSA (TPS)

Source: own study.

Table 2. Results of market model estimation for Bogdanka company

LS		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.358$	0.107	0.089	1.202	0.231	–0.068	0.281
$\hat{\beta}$	$N = 275$	0.676	0.055	12.345	0.000	0.568	0.784
LTS		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.377$	0.010	0.081	0.127	0.899	–0.149	0.169
$\hat{\beta}$	$N = 267$	0.632	0.050	12.663	0.000	0.534	0.730
$QR_{0.01}$		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.704$	–1.046	0.128	–8.181	0.000	–1.298	–0.794
$\hat{\beta}$	$N = 274$	0.606	0.024	25.431	0.000	0.559	0.653

Source: own study.

Table 3. Results of market model estimation for PGNIG company

LS		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.221$	0.044	0.096	0.461	0.645	–0.144	0.233
$\hat{\beta}$	$N = 275$	0.521	0.059	8.809	0.000	0.405	0.638
LTS		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.264$	0.065	0.080	0.809	0.419	–0.093	0.222
$\hat{\beta}$	$N = 257$	0.483	0.051	9.562	0.000	0.384	0.583
$QR_{0.01}$		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.386$	–3.042	0.127	–24.007	0.000	–3.291	–2.793
$\hat{\beta}$	$N = 274$	0.309	0.024	13.076	0.000	0.262	0.355

Source: own study.

Table 4. Results of market model estimation for TAURONPE company

LS		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.433$	–0.012	0.086	–0.143	0.886	–0.181	0.156
$\hat{\beta}$	$N = 275$	0.764	0.053	14.439	0.000	0.660	0.868
LTS		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.509$	–0.082	0.075	–1.100	0.272	–0.229	0.065
$\hat{\beta}$	$N = 266$	0.776	0.047	16.542	0.000	0.683	0.868
$QR_{0.01}$		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.713$	1.281	0.287	4.468	0.000	0.717	1.846
$\hat{\beta}$	$N = 274$	1.390	0.053	26.024	0.000	1.285	1.496

Source: own study.

Table 5. Results of market model estimation for TPSA company

LS		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.172$	0.070	0.086	0.818	0.414	–0.099	0.239
$\hat{\beta}$	$N = 275$	0.399	0.053	7.535	0.000	0.295	0.503
LTS		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.213$	0.057	0.082	0.690	0.491	–0.105	0.219
$\hat{\beta}$	$N = 270$	0.441	0.052	8.523	0.000	0.339	0.543
$QR_{0.01}$		Coefficients	Standard error	*t* Stat	*p*-value	Bottom 95%	Upper 95%
$\hat{\alpha}$	$R^2 = 0.794$	1.791	0.190	9.450	0.000	1.418	2.164
$\hat{\beta}$	$N = 274$	1.144	0.035	32.396	0.000	1.074	1.213

Source: own study.

In the Tables we present estimated parameters for three regression models calibrated for analysed time intervals. Standard error was also given. Statistical inference for determined models includes drawing conclusions on significance of parameters $\hat{\beta}$ and $\hat{\alpha}$ using student's *t*-test with known significance level applied for the test. Also provided were expected parameter estimates for all models at confidence level 0.95. Coefficients of classic LS regression should be interpreted together with test results given in Table 1. Quantile regression was formulated for substantially low quantile, because such value is taken, when for description of market behaviour we use additionally VaR (Value-at-Risk), then results of model's goodness of fit are best. In our four companies for two of them: Bogdanka and PGING all models give very similar results in estimating value of parameters and standards error of this estimation. But for next two: TAURONE and TPSA we observe that quantile regression gave differ results cause by value of outliers. We can forecast in advance this results by analysing LMS map (Figure 3).

Conclusions

The focus in this paper has been on what appears to be an unappreciated problem in empirical study, namely, a situation in which the distribution of regression residuals is not normal with fat tails. In this circumstance we clearly have ominous implications for least-squares estimation. The "corrective" proposed in this study has been the use of quantile regression (QR) which is an increasingly used robust regression procedure that corresponds to estimation by minimizing the sum of absolute errors at particular quantiles on the distribution of a model's residuals. The second appropriate method has been occurs trimmed least squares regression (LTS). This method is far more efficient than OLS, but separates only extreme not all type of outliers observations. Chosen procedures have been applied to estimation of Sharpe model which was focused on biggest companies and its benchmark from Warsaw stock exchange. The estimated regression coefficients and *t*-values was used for comparing all estimated models. To use in this circumstance estimation that is more robust than least squares seems mandatory.

Notes

[1] Trzpiot (2008).

[2] Sharpe (1971).

[3] Where *k* index was skipped.

[4] Rao (1973).

[5] Trzpiot, Majewska (2009; 2010).

[6] Fama (1965).

[7] Pratez (1972).

[8] Kon (1984).

[9] Roll (1988).

[10] Sharpe (1971); Cornell, Dietrich (1978).

[11] *Robust regression.*

[12] Ruppert, Carroll (1980).

[13] Welsh (1987).

[14] Koenker, Bassett (1978); Koenker (1982).

[15] For instance, for $\theta = 1/2$ then median of residuals from fitted model is zero: half of values from sample above the line, and half from beneath the line.

[16] For continuous random variable Z with distribution function F, it is the θ order quantile, ξ_θ is a value producing $F(\xi_\theta) = \theta$.

[17] Trzpiot (2011).

[18] Fox (1991).

[19] Maddala (2006), p. 125.

[20] Rousseeuw, Leroy (2003).

[21] $H = X(X^TX)^{-1}X^T$.

[22] Trzpiot (2007; 2008).

References

Cornell, B. & Dietrich, J.K. (1978). Mean-Absolute-Deviation versus Least-Squares Regression Estimation of Beta Coefficients. *Journal of Financial and Quantitative Analysis*, 13, 123–131.

Fama, E. (1965). The Behavior of Stock Prices. *Journal of Business*, 38, 34–105.

Fox, J. (1991). *Regression diagnostics*. Newbury Park, C.A. Sage.

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

Koenker, R. (1982). Robust Methods in Econometrics. *Econometric Reviews*, 1, 213–255.

Koenker, R. & Bassett G. (1978). Regression Quantiles. *Econometrica*, 46, 33–50.

Kon, S. (1984). Models of Stock Returns – A Comparison. *Journal of Finance*, 39, 147–165.

Maddala, G.S. (2006), *Ekonometria*, Warszawa: Wydawnictwo Naukowe PWN.

Praetz, P. (1972). The Distribution of Share Price Changes. *Journal of Business*, 45, 49–55.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* New York: John Wiley.

Roll, R. (1988). R2. *Journal of Finance*, 43, 541–566.

Rousseeuw, P.J. & Leroy, A.M. (2003). *Robust Regression and Outlier Detection*, New York: John Wiley.

Ruppert, D. & Carroll, R. (1980). Trimmed Least Squares Estimation in the Linear Model. *Journal of the American Statistical Association*, 75, 828–838.

Sharpe, W. (1971). Mean-Absolute Deviation Characteristic Lines for Securities and Portfolios. *Management Science*, 18 B1–B13.

Trzpiot, G. (2011). Wybrane odporne metody estymacji beta. *Studia Ekonomiczne 96*, Uniwersytet Ekonomiczny w Katowicach, „Modelowanie preferencji a ryzyko '11", 133–148.

Trzpiot, G., (2008). Implementation of quantile regression methodology into VaR estimation. *Studies and Papers* No. 9, University of Szczecin, 316–323.

Trzpiot, G., (2007). Quantile regression and VaR estimation. *Scientific Papers of Wroclaw University of Economics*, 1176, 465–471.

Trzpiot, G. & Majewska, J. (2010). Estimation of Value at Risk: Extreme value and robust approaches. *Operation Research and Decisions*, Vol. 20, No. 1, Wrocław, 131–143.

Trzpiot, G. & Majewska, J. (2009). Sensitivity analysis of some robust estimators of volatility. *Economics Studies* 53, 91–108, Scientific Papers of Katowice Academy of Economics.

Welsh, A. (1987). The Trimmed Mean in the Linear Model. *Annals of Statistics*, 15, 20–36.