Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

MASTER THESIS

Statistical properties of the liquidity and its influence on the volatility prediction

Author: **Bc. David Brandejs**

Supervisor: **PhDr. Ladislav Kristoufek Ph.D.**

Academic Year: **2015/2016**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, May 13, 2016

Signature

Acknowledgments

The author is grateful especially to PhDr. Ladislav Krištoufek Ph.D.and several colleagues for their comments. The usual caveat applies.

Abstract

Keywords	liquidity, risk, volatility, expected return, magic triangle, price jumps, realized variance, bi-power variation, three-stage least squares model, logit, high-frequency data, S&P 100

Author's e-mail	`david.brandejs@seznam.cz`
Supervisor's e-mail	`kristoufek@ies-prague.org`

Contents

List of Tables

List of Figures

Master Thesis Proposal

Author	Bc. David Brandejs
Supervisor	PhDr. Ladislav Kristoufek Ph.D.
Proposed topic	Statistical properties of the liquidity and its influence on the volatility prediction

Topic characteristics During the last 40 years econometric literature has offered many views on the issues concerning the measuring of the volatility (risk), volatility modelling and volatility prediction. However, there is a hole in research, which would elaborate on statistical properties of the liquidity, liquidity's dynamics and its influence on the volatility prediction. Besides volatility is further influencing returns of the assets, as it is well established empirically (higher returns are significantly correlated with higher risk). Thus, if liquidity's dynamics would be properly described, it could be rather easy to understand correctly the transmission mechanism of the whole "magical triangle", i.e. relation between liquidity, return and risk.

Volatility in financial markets is essential for asset pricing. Recent studies show that discontinuous price jumps are indeed important and have a significant effect on volatility and therefore also on asset pricing, etc. In this thesis we are going to examine whether liquidity has the influence on the price jumps. We will be using the high-frequency data.

For the purposes of monetary policy, appropriate supervisory review of the banks and lowering the effect of a major liquidity shock, precise estimates of the future (il) liquidity are of principal importance. For instance, illiquidity, rather than poor asset quality, is the immediate cause of most bank failures. Recently the topic has attracted a lot of attention of economists and e.g. Basel III introduces two liquidity standards, which should be implemented until 2018.

Hypotheses Hypothesis #1: The lower liquidity suggests higher realized risk and return

Hypothesis #2: The lower liquidity signifies higher frequency of jumps
Hypothesis #3: The deviation from standard return-risk relation are caused by liquidity

Methodology The most important part of the work will be the detection of liquidity properties. We may find proper estimates of the liquidity from high-frequency data. The core paper, which will be used in master thesis (Goyenko 2009), performs horseraces of annual and monthly estimates of each measure against liquidity benchmarks and suggests that new effective/realized spread measure will be used in most cases.

The "magical triangle" represents the instrument for effective choice among various investment opportunities, i.e. it is finding the best employment for free financial resources, as shares, bonds, commodities, mutual funds, term deposits or saving accounts. Magical triangle connects three basic components of all investments, i.e. return, risk and liquidity. High-frequency data cover intraday (tick-by-tick) transaction level data on prices, quotes, volume, order book, etc.

We will use realized variance (sum of squared returns) instead of the unobservable quadratic variation, which consists of a term representing the continuous price path and a term representing the within-day jumps. Using high-frequency data ensure that realized variance converges in probability to quadratic variation (Andersen 2003). Significant jumps in volatility may occur between the opening price of one day and closing price of the previous day. To separate these jumps we will use bi-power variation. Realized bi-power variation depends on the sum of products of absolute values of consequent intra-day returns and it can be shown that this variation converges in probability to the continuous price path component of the quadratic variation (Barndorff-Nielsen 2004). Given the mentioned properties, it is possible to estimate the price jumps as the difference between the realized variance and the bi-power variation. In the following parts of the work, standard econometric methods will be used to perform forecasting exercise in order to test our hypotheses.

Outline

1. Introduction
2. Theory concerning the Magical triangle
3. Literature overview
4. Liquidity - methodology

5. Methodology of realized variance and bi-power variation
6. Data description
7. Model for estimation the influence of volatility on liquidity
8. Results extracted from the model
9. Conclusion

Core bibliography

1. Bjorvatn, K. & C. Eckel (2006): "Policy Competition for Foreign Direct Investment Between Asymmetric Countries." *European Economic Review* **50(7)**: pp. 1891–1907.

2. Blomstrom, M. & A. Kokko (2003): "The Economics of Foreign Direct Investment Incentives." *NBER Working Papers 9489*, National Bureau of Economic Research, Inc.

3. Görg, H. & E. Strobl (2001): "Multinational Companies and Productivity Spillovers: A Meta-analysis." *Economic Journal* **111(475)**: pp. F723–39.

4. Haaparanta, P. (1996): "Competition for Foreign Direct Investment." *Journal of Public Economics* **63(1)**: pp. 141–53.

5. Haufler, A. & I. Wooton (2006): "The Effects of Regional Tax and Subsidy Coordination on Foreign Direct Investment." *European Economic Review* **50(2)**: pp. 285–305.

6. Stanley, T. D. (2001): "Wheat from Chaff: Meta-analysis as Quantitative Literature Review." *Journal of Economic Perspectives* **15(3)**: pp. 131–150.

7. Wells, L. T., N. Allen, J. Morisset, & N. Pirnia (2001): *"Using Tax Incentives to Compete for Foreign Investment: Are They Worth the Cost?"* Washington, DC: FIAS.

Author

Supervisor

Chapter 1

Introduction

During the last 40 years econometric literature has offered many views on the issues concerning the measuring of the volatility (risk), volatility modelling and volatility prediction. However, there is a hole in research, which would elaborate statistical properties of the liquidity, liquidity's dynamics and its influence on the volatility prediction. Besides, volatility further influences assets' returns, as it is well established empirically - higher returns are significantly correlated with higher risk (Fisher & Hall 1969; Neumann, Böbel, & Haid 1979; Cho & Kuvvet 2015). Thus, if liquidity's dynamics was properly described, it could be rather easy to understand correctly the transmission mechanism of the whole "magic triangle", i.e. relation between liquidity, return and risk.

Volatility in financial markets is essential for asset pricing. Recent studies show that discontinuous price jumps are indeed important and they have a significant effect on volatility (Andersen, Benzoni, & Lund 2002; Bates 2000; Eraker, Johannes, & Polson 2003) and therefore on the asset pricing as well. In this thesis, we are going to examine whether liquidity has the influence on the price jumps, we will be using the high-frequency data.

For the purposes of monetary policy, appropriate supervisory review of the banks, lowering the effect of major liquidity shocks and precise estimates of the future (il)liquidity are of principal importance. For instance, illiquidity, rather than poor asset quality, is the immediate cause of most bank failures (Robert Morris Associates 1988). Recently the topic has attracted a lot of economists' attention and e.g. Basel III introduces two liquidity standards[1], which should not be implemented later than in year 2018.

Efficient market hypothesis (Fama 1998), the key concept of the investment

[1] Liquidity Coverage Ratio and Net Stable Funding Ratio

theory, asserts that share prices reflect all information that are available to investors, in other words, share prices are the results of investors' consensus and therefore they are theoretically the best estimation of the future events (Allen *et al.* 2004). Substantial number of literature considering market efficiency hypothesis has been published (Evans 2006; Fama 1998; Himmelmann, Schiereck, Simpson, & Zschoche 2012). The problem of the hypothesis is that none of the models for expected returns is able to catch the complete patterns of the average returns for all time periods. The empirical records about the over- and under-reaction of the financial market are extensive (Barberis, Shleifer, & Vishny 1998; Daniel, Hirshleifer, & Subrahmanyam 1997; DeBondt & Thaler 1985).

Efficiency markets primarily count with a return and volatility, however, we might see that this concept is not sufficient. Abnormal returns are indeed present. Commonly in the literature the market efficiency hypothesis has been challenged by the theory of behaviorism, e.g. Hodnett & Heng-Hsing (2012) introduced their prospect theory[2]. We will try to explain the disturbances in the regression model of return and volatility via liquidity proxies.

Generally the higher demand for a share, the higher volume of the trades on the market and the higher number of investors signify higher liquidity, and thus it is rather hard to manipulate with the price. However, note that this statement is not axiomatic. Consider for instance Facebook's IPO - NASDAQ electronic stock market was not able to handle the huge flood of the demand, technical problems emerged and not all investor's demands were settled (Bunge & Strasburg 2012), despite the NASDAQ is the second biggest stock market on the world according to capitalization (Erbar 2014).

We will examine three hypotheses:

- Hypothesis #1: The deviations from standard return-risk relation are caused by liquidity
- Hypothesis #2: The lower liquidity suggests higher realized risk and return
- Hypothesis #3: The lower liquidity signifies higher frequency of jumps

The thesis is organized as follows. Chapter 2 provides theoretical background for magic triangle of investments. In this section, few basic statements

[2] Prospect theory describes the aversion concept against potential financial loss that claims that the investors rather prefer the possibility to avoid the losses than obtain a profit.

about return, risk and volatility and relationships between them are stipulated. In the Chapter 3, there is a summary of literature overview regarding liquidity measures, realized variance, bi-power variation and three-stage least squares model. Alongside the construction of variables is stated there. Chapter 4 provides a description of data employed in this thesis, particularly the high-frequency data from S&P100. Furthermore, one of the sections clarifies the method of selection the data appropriate for our testing. Graphs representing development of examined variable are attached as well. In the Chapter 5, the conditions, assumptions and chosen approaches to modeling and actual modeling and testing is carried out for above mentioned hypothesis, the results are shown in comparison tables and simultaneously commented. In the Chapter 6, the conclusions of the thesis are described.

Chapter 2

Magic triangle of investment

If we consider the wealth, i.e. summary of all investor's assets, unvarying at the moment, then every investor demand is determined by three basic criteria: return, risk and liquidity. These criteria represent the alpha and omega of the whole investment process.

When we evaluate the investment portfolio, we might notice a certain relationship between return, risk and liquidity. This relationship is called ”magic triangle of investment” (Chvátalová *et al.* 2013; Hitchner 2010). The target of every rational investor is to reach as high return as possible, alongside with the highest liquidity and the lowest potential risk of the financial instrument. However, it is not possible to reach all three goals simultaneously. An investor can prefer no more than two sides of the triangle. Investments in money market funds meet the conditions of the lowest risk and highest liquidity, but it is necessary to calculate on lower appreciation than it could be achieved by more risky investment. On the contrary, investments in shares do not satisfy the lowest risk, but comply with high liquidity and may bring high yield as the expected return is rather high (Black Book 1996). It is up to every individual investor to choose which side(s) he will prefer. Most of investors treat expected return as the main goal; they consider the return as a reward for undertaking the risk.

The magic triangle represents the instrument for effective choice among various investment opportunities, i.e. it finds the best employment for free financial resources like shares, bonds, commodities, mutual funds, term deposits or saving accounts.

2.1 Return

We perceive the return as a summary of all incomes that an investor receives from the investment; it is expressed in units of money. The return is closely connected with profitability. However the profitability is expressed by percentage, e.g. return on assets[1] or return on equity[2].

Return is the difference between current selling price and buying price from the past. If past buying price is higher than the current selling price, we suffered the loss and vice versa, if the past buying price is lower, we generated the profit. It is certainly necessary to compute with the costs included in buying/selling as well, alternatively even with the costs included in holding the asset or tax expenses.

In the empirical part we will assume the logarithmic price of the financial asset $p_t = \log(P_t)$ for day t and simultaneously we will assume intraday continuously compounded logarithmic return for time $t(i)$ between days $t-1$ and t in the following way:

$$r_{t,i} = 100[p_{t(i)} - p_{t(i-1)}] \tag{2.1}$$

We will suppose that the return is compounded from two parts – predictable component and unobservable shock. Predictable component rewards the market player for the risk of holding the financial instrument; it is rather easy to compute this part. Contrary to predictable component, the unobservable shock is not possible to be predicted.

Furthermore let us point out that we will consider no-arbitrage condition, even though this condition is rather theoretical concept, it is widely used in various financial models.

2.2 Risk

By risk we perceive the uncertainty that the real return (rate of return) will not be in compliance with the expected return. It is a threat of the loss, possible damage or the deviation from the expected rate of return. Therefore, we may characterize the risk as a variability of return. Analogically with the measurement of the return, in the case of risk, we may also measure historical risk (ex post), which follows the measurement of historical rate of return and

[1] $ROA = annual\ net\ income/total\ assets$

[2] $ROE = annual\ net\ income/shareholder's\ equity$

expected risk (ex ante), which follows the calculation of the expected rate of return.

The most favorite risk measurements for investors are absolute rate of variability, i.e. variance and standard deviation, nevertheless we may use relative variability rate, i.e. the coefficient of variation, as well.

In theory we divide the risk into systematic (undiversifiable, aggregate) and unsystematic (diversifiable, unique, specific, residual).

Systematic risk emerges from the economic system. Its sources are factors and impacts, which affect all instruments traded on the financial market. If we invest only in one economic system, then it is not possible to reduce this risk by diversification and thus, for every investment decision it is necessary to take the risk into consideration. Systematic risk is connected with political, social or economic events and with the changes in investor's preferences. There are several types of risk factors and sources of the systematic risk as the interest rate risk, economic risk, political and market risk, inflationary risk or event risk.

Unsystematic risk is related to certain instrument and it is unique for that instrument. It does not emerge from an economic system as a whole or from the financial market. It is possible to fully diminish this type of risk due to a proper choice of financial instruments. In the literature, there is no consensus among economists to create any uniform description of these types of risk. Unsystematic risk is divided into bankruptcy risk, business risk, operational risk, liquidity risk, operational risk and early redemption risk.

As the main purpose of this thesis is not a description of particular risks, we will not concern with the further explanation.

We will measure risk due to realized variance and bi-power variation. More detailed description including the equations and derivations of equations will be provided in subsections 3.2 and 3.3.

2.3 Liquidity

Liquidity means the ability to transform the investment instrument into the cash immediately and with minimal transaction costs. When it is possible to sell the instrument in a few minutes without a loss of its value, then we might declare the instrument as a liquid.

Liquidity is influenced by various factors. The most important ones are the type and character of a given instrument and the character of the market where the instrument is traded. We assume the financial market sufficiently liquid, when there is a large amount of participants, who close large amounts of trade (Prokopova 2003). Transaction costs on that type of market should be very low.

Cash, T-bills, government bonds and the most solvent shares (blue chips) are generally considered as the most liquid assets (Vesela 2011).

The investor prefers instruments with the highest liquidity. Investors that invest in less liquid instruments demand higher interest rates as a reward for illiquidity of its instrument.

Although there are no generally accepted patterns and methods to measure the liquidity of the instrument or the whole market as in case of return and risk, yet it is possible to assess the level of liquidity, alternatively to compare with commonly used available figures that record the volumes of trades, exchange rates and market capitalization. Higher number of closing deals and lower transaction costs are characteristic for a liquid market (Vesela 2011).

Similar to the risk, the description of the liquidity measures and their equations will be provided in following section 3, specifically subsection 3.1.

Liquidity is justifiable, just as two previous investment criteria, and it is not convenient to omit that during the choice of proper investment instrument.

The calculation of return, risk and liquidity of one investment instrument is not complicated. However, in reality the investor often holds the whole portfolio of instruments and usually invests his financial funds into various instruments. For the investor, who owns or just creates the portfolio, a return, risk and liquidity of individual instruments in portfolio is not the most important information anymore, nowadays he is primarily interested in the fact, how specifically an individual instrument affect return, risk and liquidity of the whole portfolio. Proportion of individual instruments have a significant role in the total market value of the portfolio.

Chapter 3

Methodology

This chapter presents the theory used in empirical section of the thesis. Methodology will be divided into four parts. In the first one, the key concept, i.e. liquidity measures and their characteristics, are introduced. In the second and third part the derivation of realized variance and bi-power variation and simultaneously their interaction are described. In the last part of the Methodology chapter, we will introduce three-stage least squares model.

3.1 Liquidity measures

Measurement of (il)liquidity depends on the approach the researcher is pursuing for. According to Mancini *et al.* (2013) liquidity measures can be separated into three types - price impact, trading costs and price dispersion (volatility). As we examine the impact of liquidity on volatility, we excluded the volatility measures. Furthermore because our dataset does not include the spread, we excluded the trading cost proxies (spread proxies) as well. Therefore we will focus on the price impact proxies.

The core paper for the detection of liquidity properties is Goyenko *et al.* (2009), the findings from this paper inspired this master's thesis. In this paper authors examine both previously known liquidity proxies and newly added liquidity proxies given by the authors. They intend to carry out comprehensive comparison of the proxies, such that the investors shall have an instrument to assess whether their trading strategies are truly profitable regarding the cost of trading. The authors perform horseraces of annual and monthly estimates of each measure against liquidity benchmarks and suggest, that new effective/realized spread[1] measure should be used in most cases as spread proxies and Amihud Illiquidity measure as the price impact proxy.

Furthermore, the authors asserted that proper estimates of monthly and annual liquidity shall be derived from low-frequency (daily) stock returns, this assertion allows us to study liquidity over longer time periods (tens of years) and across various countries.

According to Goyenko *et al.* (2009) number of papers suggested liquidity measures based on daily returns, but these measures rarely test whether the measures are connected with actual transaction costs. The authors discover, that low-frequency measures (monthly and annually) can capture high-frequency measures of transaction costs, it implies, that an endeavor of using high-frequency data does not worth the costs. The elaboration of this phenomena could assess whether investor's strategies are really profitable (cost of trading included).

Goyenko *et al.* (2009) claim that the liquidity proxies are not often tested whether they truly capture the transaction costs due to limited availability of actual trading costs. However, this issue concerned the data in US markets before 1983 and as we will use the dataset only for years 2013 and 2014(for the data description used refer to subsection 4.2), we further do not have to deal with the issue.

Price impact proxies measure impact of a trade on the quoted price. The volume of the impact depends on the volume of a trade, i.e. higher volume of a trade signifies higher liquidity of the market and lower price impact.

Goyenko *et al.* (2009) suggest three low-frequency price impact proxies that prevail the remaining ones, these are Amihud Illiquidity, Amivest Liquidity and Roll proxy and will be described in the following subsections.

3.1.1 Amihud Illiquidity

Notion that low price impact is related to high liquidity is captured in the following equation presented by Amihud (2002):

$$Amihud\ Illiquidity = Average\left(\frac{|r_t|}{volume_t}\right) \tag{3.1}$$

The basic idea behind the formula is very trivial - absolute value of return

[1] Effective spread means the spread that is actually paid by investors, Realized spread is the difference between average bids and offers over a certain period of time. Small spread signifies that the liquidity of given instrument is high.

for day t is divided by traded volume within given day t for all positive-volume-trading days, since the fraction is not defined for zero-volume days. The Amihud Illiquidity is then calculated as the average of individual day fractions over the whole examined period. Low Amihud Illiquidity indicates high level of liquidity and lower price impact. Amihud (2002) states that Equation 3.1 captures "Daily price response associated with one dollar of trading volume".

In the empirical part we will use slightly adjusted Amihud Illiquidity measure based on high-frequency data:

$$Amihud_t = \frac{1}{M} \sum_{i=1}^{M} \frac{|r_i|}{volume_i}, \tag{3.2}$$

where r_i is an intraday return, $volume_i$ is a relevant trade volume to the intraday return and there is M intraday returns within given day t.

3.1.2 Amivest Liquidity

Similar to Amihud Illiquidity, there is another price impact measure - Amivest Liquidity. This measure was developed by Cooper *et al.* (1985). Basically it is just an inverse function of Equation 3.1:

$$Amivest\ Liquidity = Average\left(\frac{volume_t}{|r_t|}\right) \tag{3.3}$$

Traded volume within day t is divided by absolute value of return for all non-zero return days. The Amivest Liquidity is then again calculated as the average of individual day fractions over the whole examined period. Low Amivest Liquidity measure suggests low level of liquidity and higher price impact.

In the empirical part, we will again use the high-frequency modification:

$$Amivest_t = \frac{1}{M} \sum_{i=1}^{M} \frac{volume_i}{|r_i|}, \tag{3.4}$$

3.1.3 Roll

The last liquidity measure used in this diploma thesis was introduced by Roll (1984). It estimates the effective bid-ask spread on the basis of serial covariance between consecutive price changes. To obtain the unbiased results while using Roll proxy, few conditions have to be fulfilled - the stock exchange has to be informationally efficient market, distribution of price changes is supposed to be stationary and the true value of the share has to be exactly equal to average of bid and ask prices. We assume these conditions to be fulfilled as the shares in our dataset are from S&P100 index, i.e. these are the shares of largest US companies from the most developed stock exchange markets.

Assuming the three above mentioned conditions, Roll (1984) proved that $Cov(r_i, r_{i-1}) = -s^2/4$, where s is bid-ask spread. It means that if no new information occurs, price changes only from ask to bid and bid to ask price, depending on the previous trade (whether it was buy or sell). Goyenko *et al.* (2009) stated that if $Cov(r_i, r_{i-1})$ is non-negative, i.e. Roll is not defined, the Roll should be 0. Therefore, the equation is:

$$Roll_t = \begin{cases} \sqrt{-4Cov(r_i, r_{i-1})} & \text{if } Cov(r_i, r_{i-1}) < 0 \\ 0 & \text{if } Cov(r_i, r_{i-1}) \geq 0 \end{cases} \tag{3.5}$$

Roll estimator is liquidity spread proxy according to Goyenko *et al.* (2009), and as stated above, we do not have a data with bid and ask prices. However, we have made an assumption that the true value of the stock is the average of bid and ask price and we have assumed that this average price is the adjusted closing price from our dataset, for further explanation of adjusted closing prices in the dataset refer to subsection 4.2.

Because Roll is the spread proxy, the lower Roll signifies the higher volume of liquidity. We would like to point out that the period with high liquidity, i.e. with the $Roll = 0$, signifies that bid-ask spread is also 0 and it means that the trading costs are 0 as well. Obviously, it is not truth in real life. Even the most liquid shares in NYSE or NASDAQ have a gap between bid and ask prices.

Besides, Novak (2012) warns that it is not possible to assume any "true value" in the time when an uncertainty of the final price persists. And as stock markets do not calculate with the end of the market like e.g. prediction markets, Novak (2012) states that there is nothing like objectively correct determination of the price given available information. Simultaneously it is not possible to find out, whether the market is efficient, because market efficiency is dependent on the correct valuation of the assets in that exact moment.

In the empirical part we have employed Amihud Illiquidity, Amivest Liquidity and Roll proxies based on high-frequency data, even though Goyenko *et al.* (2009) asserts that it does not worth the costs for the investors' prediction. However, as the topic of the master's thesis is the properties of liquidity and as we do have an access to high-frequency data, we have decided to use them to obtain as precise results as possible.

All three proxies provide statistically significant and useful measures, despite the fact they are computationally easier ones. Some other measures were introduced by Goyenko *et al.* (2009), but these measures will not be employed in the empirical part as it is possible to reach the same results by these computationally less complicated models. In fact, according to authors, the Amihud Illiquidity shows the best results overall. This liquidity measure of price impact proxies dominates the other measurements that are often used in literature like Pastor and Stambaugh's Gamma (Pastor & Stambaugh 2003) or extended Amihud proxies (Goyenko *et al.* 2009).

3.2 Realized variance

We will use realized variance (sum of squared returns) instead of the unobservable quadratic variation. Quadratic variation consists of a term representing the continuous price path and a term representing the within-day jumps. Using high-frequency data ensures that realized variance converges in probability to quadratic variation (Andersen *et al.* 2000).

We will employ realized variance measures instead of widely used variable models like (G)ARCH or Stochastic volatility, because these traditional latent variable models are not able to capture conditional volatility, which is highly persistent. Besides, parametric models do not employ high-frequency data, multivariate extensions are difficult and standardized returns are not Gaussian, thus forecasts are not accurate.

Meanwhile realized variance measures are based on high-frequency data, they do not depend on any parametric assumptions and high-dimensional multivariate modeling is possible. However, the main advantage over parametric models is, that it is feasible to capture the volatility (Barunik 2013b).

In finance, volatility stands for the variation of value of the asset, thus we may count it as a function of returns. We will not assume simple returns, in the subsection 2.1 we have already stipulated the continuously compounded return by the Equation 2.1.

Let us suppose, as Barunik (2013a) suggests for high-frequency data, that continuous-time stock price follows a jump-diffusion model and contain noise, therefore:

$$y_t = p_t + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2), \tag{3.6}$$

$$dp_t = \mu_t \ dt + \sigma_t \ dW_t + c_t \ dJ_t, \tag{3.7}$$

where W_t stands for independent Brownian motion, J_t represents constant-intensity Poisson process and jump magnitude is controlled by $c_t \sim N(0, \sigma^2)$. These equations indicate that is is possible to separate noise and jumps from the pricing model and implicitly reveal the "true" process generating the data. We will separate daily realized variance into the continuous (persistent) and jump components in the way Barndorff-Nielsen & Shephard (2004); Huang & Tauchen (2005) suggested.

Let us consider a day t with return r_t, then the equation for daily realized variance is expressed as a sum of squared intraday returns:

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2 \qquad \text{for } t = 1, .., T\text{days}, \tag{3.8}$$

where M stands for the number of intraday returns per day t. We will employ high-frequency data, therefore we have to consider the influence of the market microstructure noise as it might bias the approximation of the variance. We will discuss this matter in the subsection 4.1.

Moreover, according to jump-diffusion process in 3.7, we can construct the quadratic variation of jump-diffusion process:

$$QV_t = \int_0^t \sigma_s^2 \, \mathrm{d}s + \sum_{j=1}^{t} J_s^2 \tag{3.9}$$

In 3.9 we can see an integrated variance part and the variation of jumps and as Andersen *et al.* (2000) proved, an asymptotic realized variance goes to quadratic variation:

$$RV_t \rightarrow \int_0^t \sigma_s^2 \, \mathrm{d}s \tag{3.10}$$

Therefore, we may conclude, that realized variance is unbiased and consistent estimator of integrated variance, if:

$$S \frac{r_t}{\sqrt{RV_t}} \sim N(0, 1) \tag{3.11}$$

Realized volatility is equal to square root of realized variance.

3.3 Bi-power variation

Jumps in volatility have been engaged by many financial economists in recent years as their importance has been quantified by high-frequency data. Literature (Barndorff-Nielsen & Shephard 2004; Mykland, Shephard, & Sheppard 2010) asserts, that the jumps are quite common in the current price development.

Andersen, Benzoni, & Lund (2002); Bates (2000); Eraker, Johannes, & Polson (2003) came up with the proof that discrete jumps may occur in asset prices. This finding would make the price path discontinuous. However, Andersen, Benzoni, & Lund (2002); Bates (2000); Eraker, Johannes, & Polson (2003) also proved that the jumps are present rather occasionally and there is no consensus concerning the distribution of the volume of the jumps.

Significant jumps in volatility may occur between the opening price of the current day and the closing price of the previous day. To separate these jumps it is necessary to use bi-power variation. Realized bi-power variation depends on the sum of absolute values of consequent intra-day returns and it can be shown that this variation converges in probability to the continuous price path component of the quadratic variation. It is possible to estimate the price jumps as the difference between the realized variance and the bi-power variation (Liptak 2012).

Barndorff-Nielsen & Shephard (2004) introduced the realized bi-power variation as an extension to realized power variation. Bi-power variation shows better results in the estimation of integrated variance in stochastic volatility models. Authors rewrote the returns as $\Delta X_{t_i} = X_{t_i} - X_{t_{i-1}}$, assuming that $0 = t_0, t_1, \ldots, t_n = 1$, and adjusted realized variance as follows:

$$RV = \sum_{0<t_i\leq 1} (\Delta X_{t_i})^2 \tag{3.12}$$

On the basis of 3.12, Barndorff-Nielsen & Shephard (2004) defined the bi-power variation as:

$$BV_t = \frac{\pi}{2}\frac{M}{M-1}\sum_{i=2}^{M} |r_{t,i-1}||r_{t,i}|, \tag{3.13}$$

Barndorff-Nielsen & Shephard (2004) proved that as $M \to \infty$, BV_t converges to the daily integrated variance that is not affected by the jump. Altogether it means that the result of $RV_t - BV_t$ is a consistent measure for the jump component in the total daily realized variance. The test statistic and additional calculation will be provided in the empirical part.

Bi-power volatility is equal to square root of bi-power variation.

3.4 Three-stage least squares model

We plan to use the three-stage least squares model introduced by Zellner & Theil (1962) in the empirical part of the master's thesis.

Using the simultaneous equations arises the issues of identification and completeness. The system of structural equations is complete if there are as many dependent variables as the number of equations. Identification problem means that the equations are not identified and thus, every linear combination of these equations looks exactly the same.

Let us assume that the structural equation has a left-hand side dependent variable y_1. Furthermore, g_1 is the number of right-hand side dependent variables, k_1 is the number of right-hand side independent variables and at he same time the right-hand side dependent variables g_1 are correlated with disturbance ϵ_1, then we might say that OLS estimation is biased and inconsistent.

A necessary, but not sufficient, condition for identification of the system says that the number of excluded independent variables from the equation is bigger or equal than the number of right-hand side included dependent variables. Thus, this order condition demands: $k_2 \geq g_1$, where $k_2 = K - k_1$ and K is the number of all independent variables from structural equations.

We may derive the degree of identification as $I = k_2 - g_1$ and determine that the system is under-, just and over-identified, when $I <=> 0$.

To solve the identification problem, besides the order (necessary) condition, the rank (sufficient) condition has to be fulfilled, as well.

Consider the system of equations that contains 5 exogenous variables and 2 endogenous variables. In the first equation only one exogenous variable is included, therefore the order condition $K - k_1 \geq g_1$ (in this case $4 \geq 1$) is satisfied with inequality. The second equation contains 4 exogenous variables and none of the endogenous variables (on the right-hand side). Thus the order condition is satisfied again with inequality $(4 \geq 0)$.

The rank condition is $rank(A\phi) = G - 1$, where

$$\mathbf{A} = \begin{pmatrix} 1 & -\beta_{12} & -\gamma_{12} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -\gamma_{23} & -\gamma_{24} & -\gamma_{25} & -\gamma_{26} \end{pmatrix}$$

and ϕ_1 and ϕ_2 for the first and the second equation, respectively:

$$\phi_{\mathbf{1}}^{\mathbf{T}} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$\phi_{\mathbf{2}}^{\mathbf{T}} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

After multiplying A with the corresponding ϕ we get:

$$\mathbf{A}\boldsymbol{\phi_1} = \begin{pmatrix} 0 \\ -\gamma_{23} - \gamma_{24} - \gamma_{25} - \gamma_{26} \end{pmatrix}$$

$$\mathbf{A}\boldsymbol{\phi_2} = \begin{pmatrix} 1 - \gamma_{12} \\ 0 \end{pmatrix}$$

Thus, the rank condition for the first equation is satisfied if $-\gamma_{23}-\gamma_{24}-\gamma_{25}-\gamma_{26} \neq 0$ and the rank condition for the second equation is satisfied if $-\gamma_{12} \neq 1$. If these conditions are satisfied, then both equations are over-identified.

Chapter 4

Data

Data are crucial part of every empirical paper as the results directly depend on them. This section will provide the justification of the selection of particular dataset, because we find necessary that the reader gets an overview of the dataset, so he will better understand the calculations that will follow. First subsection introduces properties of high-frequency data, second one describes the selected dataset and the third one states the basic information about S&P 100.

4.1 High-frequency data

Involving high-frequency data into econometrics model is an issue that emerged at the turn of millennium. It is possible to record high-frequency data due to technological progress. HF data are widely used both by econometric researchers and market traders as they tend to have as much precise data as possible. It is further driven by growing importance of intraday trading.

The crucial task in high-frequency data selection is choosing the proper sampling frequency to avoid market microstructure noise, e.g. bid-ask spread, late reporting, price discreteness, rounding errors or screen fighting. Microstructure noise is a disturbance that makes high-frequency estimates unstable.

There is not an absolute compliance among economists in the selection of the appropriate sampling frequency for high-frequency data. Various analysis demand various sampling frequencies. Ryu (2011) suggests 1 min resolution, Pooter *et al.* (2008) suggest 30 minutes resolution; we selected 5 minutes resolution that is most often proposed by the literature (Goyenko *et al.* 2009;

Andersen *et al.* 2000). However, the decision about the sampling frequency selection will be always partially arbitrary.

As Hasbrouck (2009) proposed, we have chosen our data according to five criteria that stock has to meet to get unbiased results:

- common stock
- stock has to be present on stock exchange during the whole examined sample
- primarily listed on NYSE, AMEX or NASDAQ
- stock does not change primary stock exchange, tick, symbol or CUSIP[1] over the year
- listed in CRSP[2]

On the basis of these criteria we had to exclude Google shares from our dataset, because 27th March 2014 Google created an entirely new class of stocks and issued them to stockholders as a stock dividend instead of more usual simple doubling the number of existing shares (with the new share amounting half of the previous value). It means that in our dataset since 1st January 2013 until 31th December 2014 there are only 98 shares, instead of the whole bunch of 100 shares traded on S&P100. For more information about the used dataset refer to the following section[3].

4.2 Selection of the data

We have employed high-frequency data provided by vendor QuantQuote.com, this dataset has been recommended by Caltech (2015), the authors ran qual-

[1] CUSIP - code that identifies a financial security

[2] Center for Research in Security Prices

[3] Note that we had to take into a consideration the numeric imprecision in Microsoft Excel as we have used the Visual Basic for Applications (VBA) to process the high-frequency data. Basically, Excel is limited to an arbitrary number of significant digits. We have used the logarithm in every calculation. Logarithm is very likely to have many significant digits and when there is a multiplication between two logarithms, or even a chain-multiplication between more than two logarithms, the number of significant digits might grow even exponentially. Thus the result would have had much more significant digits than the "initial logarithms". The result of each of those operations is rounded at an arbitrary decimal place. After a hundred or so operations the cumulative error starts creeping up. We have checked that our data are accurate at least for five decimal places, it should be enough to obtain accurate results.

ity screening over dozen vendors from all US national/regional exchanges and QuantQuote.com turned to prove the best results.

We have chosen the two-year panel dataset consisted of 100 publicly traded shares on S&P100 as at 15th May 2015 in date range since 1st January 2013 to 31st December 2014. For more detail about the particular items of S&P100 used in this master's thesis refer to Appendix A. Due to excluding Google shares from the dataset, as it was described above, we have 98 shares quoting during 504 active trading days (there is no trading activity in the stock exchange during the weekends and US stock market holidays), in total it gives us 49 392 data day-points. The dataset we have obtained includes date and time of the trades during the trading day (in 5 min sampling frequency as described in subsection 4.1), open/high/low/close price, volume, splits, earnings and dividends. QuantQuote.com automatically provides closing prices as the adjusted ones. The concept of adjusted prices is explained further.

The closing price signifies, as the term suggests, the price that is quoted at the very end of the time range, i.e. in our case at the very end of 5 min sampling frequency. However, the price of the share is affected not only by the markets' interaction of supply and demand. It may happen that the company decides to pay a cash dividend, then the stock price decreases about the amount of the announced divided to protect itself against traders that would want to hold the share only for one day to collect the dividend. Any other distribution to shareholders as stock dividends, stock splits, earnings announcement or rights offering would influence the price of the share as well.

To compare a stock's performance over a period of time the adjusted closing price has been established. The most common computation of adjusted prices comprises from two types of adjustments - stock splits and dividends. Stock splits are rather trivial to calculate as the previous price is solely divided into multiple shares, however intrinsic value of the company remains the same. Let us provide an example. On Monday, Share A closes at $50 per share, on Tuesday two-for-one stock split becomes valid, Share A opens at $25 and closes at $26, therefore intraday return is $1. Considering only closing prices we would have to conclude the decrease about $24 (50 - 26). But if we assume adjusted closing price, the adjusted closing price for Monday changes to $25 and adjusted closing price for Tuesday of $26 indicates that in fact there is a gain $1.

Calculation of cash dividends' adjustments is also straightforward. The amount of the dividend is subtracted from the previous closing price. Again let us show the example. Closing price of Company ABC equity is \$50 per share on Monday, Company ABC decides to distribute dividends after the closing time of the stock market, the dividend will be amounted \$2 per share, thus the adjusted closing price will be equal to \$48. However, then it may happen that if the dividend is larger than the closing price of the share, than the adjusted closing price at the end of the day may become negative. Intuitively it makes no sense, thus it has to be taken care of.

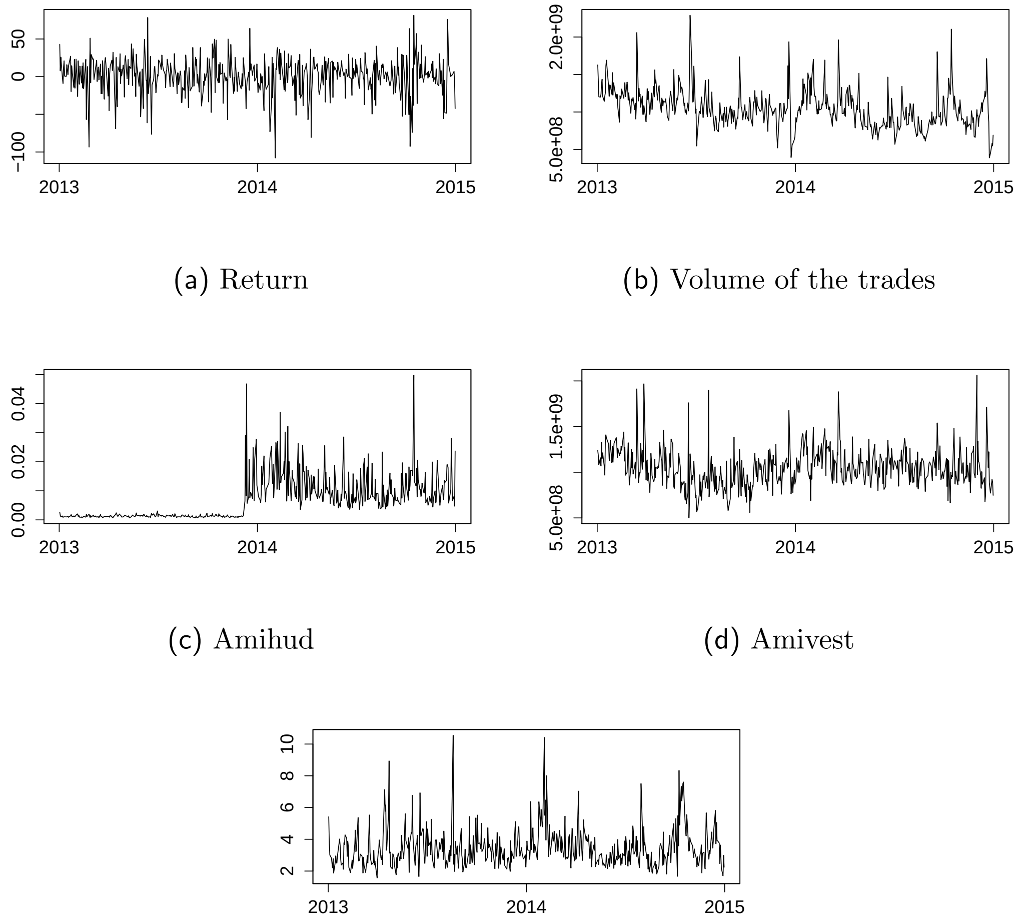
A 2:1 stock dividend signifies that the shareholder will receive three additional stocks for every single stock that he already possesses. The calculation of the adjusted closing price for the Company ABC from the previous paragraph will be following: $\$50 * (1/2) = \$25$. The computation of other corporate actions such as rights offering is more complicated, however as mentioned above QuantQuote.com already took care of the adjusted closing prices and provided them in the dataset by default, thus we did not have to deal with the issue.

We have attached the original data provided by QuantQuote.com, refer to Appendix C. Please note that in each file there are as many rows as was the number of 5 min sampling frequencies with non-zero trades in the day for the particular stock.

We have not established any minimal M, i.e. number of 5 min resolutions in the day, because if we had done so, we would artificially create a subsample and we could omit a significant observation. Minimum M is 42, it means that the smallest data point has 42 5-min resolutions included, maximum M is 155, average M is 92 and median 86.

Further note that the closing price, more precisely adjusted closing price, is not the same as the opening price in the following 5 min sampling frequency. Every sampling frequency is taken considered unique, it means that the value of the opening price is the value of the first trade carried out in the 5 min sampling frequency, not the value of the last trade in the previous 5 min frequency. The closing price is, as regularly, the price of the last trade within given 5 min resolution.

Let us describe the data graphically. In the Figure 4.1, we provide the development of the daily return and total volume of the daily trades, as we presume that for these variables the absolute value has a corresponding value. Moreover, we attach the graphs for three liquidity proxies as the liquidity is a primarily examined variable. Each day is compounded from the sum of all 98 shares from our dataset. The development concerns the period since 1 January 2013 until 31st December 2014.



(a) Return

(b) Volume of the trades

(c) Amihud

(d) Amivest

(e) Roll

Figure 4.1: Development of the variables since 1/1/2013 until 31/12/2014

Even though the graph of return looks like that during the years 2013 and 2014, the sum of returns across all shares is negative, the opposite is truth. When an investor had bought one share for all 98 companies from our dataset at the very beginning of the year 2013 and kept the shares until the end of the year 2014, he would earn approximately $560, because it is the sum of all daily returns calculated from high-frequency data for all 98 shares in these two years.

We may also notice the interesting development of Amihud Illiquidity. Despite the rising market that increased the share prices, this liquidity proxy showed constantly low value approximately until the December 2013. It had been caused by low number of stock splits during the year 2013 Rosenberg

(2013). And because the stock splits influence the calculation of the high-frequency Amihud Illiquidity (more than the calculation of Amivest Liquidity or Roll proxy), we may notice the difference between years 2013 and 2014.

The daily average volume of total trades is 1 053 270 230 , it means the volume of 10 747 655 trades per one share in average. The "busiest day", the day with the largest volume of trades (2 290 595 069) was 21st June 2013, the lowest trading volume was on 24th December 2013 and 24th December 2014 (393 112 683 and 387 244 674). The most traded share was the Bank of America Corp with its volume of trades 51 744 164 005 for the whole two-year period, meanwhile during the same period the Simon Property Group, Inc. had the volume of trades only 660 942 185.

4.3 S&P 100

The S&P 100 Index is a stock exchange index for the United States companies that is presented by the rating agency Standard & Poor's. The index measures the performance of 100 blue-chip US companies. The index is a subset of more commonly known S&P 500, however, its average market cap is twice as big as that of the S&P 500 - specifically $142 bn for S&P 100 and $68 bn for S&P 500 in the April 2014 (Rhoads 2014).

To be included into the S&P 100, the companies need to have listed options. Index options are traded with the "OEX" ticker symbol. The index involves the options of the 100 largest and most established US companies.

The index is constructed following these criteria:

- The company is included in S&P 500
- The minimal market capitalization of the company is $4,6 bn
- At least half of the outstanding shares has to be available to be traded
- The company must have the positive earnings for the previous four quarters in total.
- The company has highly tradable shares.
- The prominent companies are chosen across the individual industries, meanwhile the sector balance is considered.

- The company's shares are traded on NYSE, NYSE Arca, NYSE MKT,REITs, NASDAQ Global Select Market, NASDAQ Global Market or NASDAQ Capital Market.

The index launches since 15th June 1983. It is calculated in EUR and USD (our dataset is in USD).

Chapter 5

Models & Results

Generally we have put a great effort into the basic analysis. Given the liquidity and volatility are not normally distributed, they are even not getting closer to this distribution or any kind of symmetry, we adjusted them due to logarithm as the literature suggests (Xue 2014).

Using the logarithm for the volatility does not cause any problem since its equation consists of squared returns, i.e. the logarithm of volatility will be always non-negative. The only issue might occur while the return would be equal to zero. Note that there are some zero-return days in the dataset. These daily returns were calculated as the summary of intraday returns, see Equation 2.1, it means that the negative and positive intraday returns are offseted and the result is a zero in summary. However, during the day, there were some movements of the returns and these movements caused an intraday volatility. And because we have calculated the daily realized variance on the basis of intraday volatility (high-frequency data have been used), we might see in our dataset that there are no daily zero-volatility datapoints, thus the input for logarithmic volatility is not only non-negative, but even positive. Naturally it implies that logarithm of realized variance exists in every datapoint.

In literature, logarithm of the liquidity causes more problems, because sometimes during the day a sampling frequency can have a zero trading volume, thus logarithm does not exist. However, QuantQuote.com provided us the dataset only with the sampling frequencies in which any market trade of given share took place, thus we have avoided the sampling frequencies with zero trading volume. We only had to adjust the calculation of Amihud Illiquidity and Amivest Liquidity (see Equations 3.2 and 3.4, respectively), because these measures include an intraday return in their calculations and the intraday return is zero in some cases. Subsequently it implies that the Amihud Illiquidity is zero, as well and the equation of Amivest Liquidity does not make sense due to dividing by zero. Therefore in these circumstances we consider the Amivest Liquidity to be zero instead.

We also adjusted four daily data for the following four different stocks - for each FedEx on 30th October 2014, Costco on 16th December 2014 and both Honeywell and Pepsi on 11th Novemeber 2014, there is a negative volume in one of 5 min sampling frequencies. It is due to the cancelation of the trade and it causes Amihud Illiquidity to be negative. Thus we had to adjust these four daily Amihud proxies based and set them zero.

We considered deleting these four shares from the dataset, however we wanted to keep the dataset strongly balanced for the testing in Stata, it means that all shares will have data for all periods (t=1,...,504). If we had deleted these four shares, the data would become unbalanced.

Obviously we did not use the logarithm for a return since returns might be zero or negative.

This is a general setup, particular modifications will be presented further when appropriate for particular calculation.

5.1 The lower liquidity suggests higher realized risk and return

In the theory the magic triangle is well described. In this empirical part we will include the liquidity and describe the disturbances in the magic triangle.

We will perceive the liquidity as an exogenous variable, it is a generally accepted assumption in the literature (Han & Lesmond 2011), and return and volatility as endogenous variables.

Our basic assumption about the model is expressed via following two regressions:

$$return_t = \alpha_0 + \alpha_1 volatility_t + \alpha_2 liquidity_t + \epsilon_{1t}, \tag{5.1}$$

$$volatility_t = \beta_0 + \beta_1 return_t + \beta_2 liquidity_t + \epsilon_{2t}, \tag{5.2}$$

Return and volatility are contained in both equations as there is an evidence of leverage effect's existence (Figlewski & Wang 2000). Leverage effect describes the relationship between stock returns and realized variance - when the price is falling, the variance is rising. We will take this phenomena strongly into consideration, because Figlewski & Wang (2000) have confirmed the leverage effect on the S&P 100 shares. We will consider the phenomena even though, the authors have found few anomalies, e.g. weaker or non-existing leverage effects, however, these anomalies have been found only for companies. The authors confirmed that for the whole market (that we care more in the thesis than a particular company) the phenomena persists.

5.1.1 Three-stage least squares model

Given the simultaneous equations and the assumption of cross-correlated error terms between the two equations, we have chosen the three-stage least squares estimator (3SLS) described in the subsection 3.4 as the model that should capture the Equations 5.1 and 5.2 in the best way and produce even better efficiency than for example 2SLS. As Oberhofer & Kmenta (1973) proved, 3SLS is asymptotically more efficient than 2SLS due to iterations to higher stages. Final estimators are then identical to maximum likelihood estimates.

It is important to realize that due to limited information, the 3SLS model is efficient just over the estimation of one equation. It means that 3SLS estimator does not produce the consistent estimates. However, in our case, it is convenient to use 3SLS method as a system estimator that uses the cross-equation restrictions on the structural equations.

The employment of the 3SLS method should provide us the answer not only on the hypothesis # 2 (whether the lower liquidity suggests higher realized risk and return), but also the on the hypothesis # 1 (whether the deviation from standard-risk relation are caused by the the liquidity).

5.1.2 Standardization

The econometric program used in the master's thesis (Stata) is not able to compute simultaneous equations for panel data. Thus, we had to transform the data into the non-panel framework. There are two possibilities to do so, first one, i.e. using explicit dummies for observational units or second one, i.e. deviate the data from means[1]. We have chosen the latter option. Note that it would be also possible to run the OLS for panel data equation-by-equation. The estimation would be consistent, but it would not be as efficient as with transformed 3SLS and when one equation would not be set properly, it would affect the second equation as well. Nevertheless, we have performed the comparison with OLS on the final model to reach the robustness check.

Since we have already subtracted the mean from the data, it only makes sense to standardize the variable alongside. Thus, each datapoint has been transformed in the following way:

$$z = \frac{x - \mu}{\sigma}, \tag{5.3}$$

where x is the particular datapoint, μ is the mean of the population, σ is the standard deviation of the population and z is the standardized value. The absolute value of z signifies the distance between the datapoint and average value of the population in units of the standard deviation. When z is negative, the value of the datapoint is below average and vice versa for positive value of z.

Naturally, standard deviation has been calculated separately for each variable and also for each share to reach more sensible estimations.

[1] Both these adjustments for panel data are allowed only for fixed effects model. Generally (not just in the econometric programs Stata or R) it is not possible to use 3SLS estimation if you want to model random effects.

5.1.3 Autocorrelation

Given time-series nature of the data, we had to take the autocorrelation into the account, because there is an expectation that the error terms in time t and s might be correlated. Therefore, the crucial task in the process of modelling is to reach a model that does not contain "strongly autocorrelated" error terms. "Strong autocorrelation" means that the time series of error terms include a unit-root. It would mean that our model will not make the sense and the analysis would have to be reconstructed. The "weak form of the autocorrelation"[2] of disturbances, as the name suggests, would not cause such a serious trouble as the strong form and it could be easily treated by heteroskedasticity and autocorrelation consistent standard errors (HAC SE) or robust standard errors, both these adjustments could be reached by build-in routine in Stata.

Note that 3SLS estimator takes care of the cross-correlated error terms between the equations, however, it does not solve the issue of the autocorrelation of error terms inside one equation.

5.1.4 Lagged variables

Our initial assumption was that the first equation should contain the lags of return that would take care of possible autocorrelation and the second equation should involve the lags of return and volatility since the volatility has a "long memory".

Because we also take into the consideration the economic sense during the modelling and we tend to avoid the simple data mining, we have decided that the maximum possible lag length shall be 5. The decision is arbitrary, however, we believe that it is reasonable, because our data are from S&P100, i.e. these are the shares of largest US companies traded on largest stock markets as NYSE or NASDAQ. Thus we might assert that any new information or e.g. exceeding stock return would be incorporated in during one trading week, i.e. 5 days, and thus, we have chosen 5 as the maximum possible lag length.

To decide which exact lag length should be chosen as the most appropriate, we have followed Koop (2014). The author asserts that the selection of the proper lag length shall be carried out due to t-test (p-value) of particular lagged variable in the model or due to information criteria. Koop (2014) suggests to include the maximum possible lag length (we have chosen the $qmax = 5$)

[2] "Weak form of autocorrelation" means that error terms for various times $t = 1, \dots, T$ are drawn from different error terms that are not exogenous variables.

that seems to be reasonable into the model and remove one-by-one the lagged variables with the lowest value of t-test. We have followed the author and repeated this procedure as many times as necessary, i.e. until all variables' p-values were less than the significance level $\alpha = 0,05$. However, since we have 2 endogenous variable and 5 exogenous variables plus there are other 5 lagged variables for both endogenous and exogenous variables and all these variable are included in both equations, the modelling did not seem to be reasonable, the results did not make any economic sense and we just tend to mine the data.

Therefore, we have decided for the second possibility suggested by Koop (2014) - use the information criteria and assess the models due to the lowest value of the given criterion.

5.1.5 Evaluation of models

To assess individual models we have used two criterions provided by Stata programme - Akaike information criterion (AIC) and Bayesian information criterion (BIC). These are defined as:

$$AIC = -2ln(likelihood) + 2k, \tag{5.4}$$

$$BIC = -2ln(likelihood) + ln(N)k, \tag{5.5}$$

where k is the number of estimated parameters and N is the number of observations. The theory behind above mentioned criterions is rather easy - the smaller the value of AIC (BIC respectively) is, the better the model fits the data.

We may derive from the equations of AIC and BIC that the information criteria penalize for excessive number of parametres.

Note that all that matters when we evaluate the model with the information criterion is the difference between AIC (BIC) values for two particular models. The actual magnitude or whether the criterion is positive or negative does not play any role. Imagine if we had changed the units of the data, the value of information criteria would change substantially, however, the difference between the values of AIC (BIC) would not change a bit.

Furthermore, note that we do not assess the model according to R-squared since it is rather problematic for simultaneous equations models. Sribney *et al.* (2015) stated that R^2 has no statistical meaning in the context of 2SLS, 3SLS or IV estimator. R^2 might be suppressed as it may even acquire negative values with 3SLS estimation. It may happen, because model sum of squares can be negative. Let us assume the formulas:

$$R^2 = MSS/TSS \tag{5.6}$$

$$MSS(model\ sum\ of\ squares) = TSS - RSS \tag{5.7}$$

$$TSS(total\ sum\ of\ squares) = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{5.8}$$

$$RSS(residual\ sum\ of\ squares) = \sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{5.9}$$

When MSS is negative, R^2 is negative, as well. MSS is negative, when RSS is larger than TSS and it happens when $\bar{y}$ is a better estimator of y than $\hat{y}$. And it may happen with 3SLS estimation (even when the constant term is included).

But it doesn't mean that the estimations are wrong when RSS is larger than TSS. Let us point out that we care about the parameters in the structural equation and if the model estimates theses parameters with a satisfactory standard errors, then we might consider it as a good model, regardless of MSS or R^2.

5.1.6 Approaches to modelling

We had estimated the results without a constant. It makes sense, because we had already adjusted the data due to standardization - the average was also subtracted. Besides, the models without a constant consistently proved to show up better results both due to AIC and particular p-values of the constant. For both equations, p-value of the constant tend to be over 0,9 for most of the cases.

Because we want to assess the model due to information criteria and we stated that both AIC and BIC penalize for excessive number of parametres, we tried to keep as few parametres as possible. Therefore, we have started our modelling with the models without any lags and included the two endogenous variables (realized variance and return) and only one liquidity proxy or the number of intraday returns (both rounded and exact).

In the Table 5.1 we provide the results of these models.

As stated in subsection 3.4, while using the 3 SLS model, the crucial task is to assure the completeness and identification of the system. Already the initial regression Equations 5.1 and 5.2 are complete as the number of dependent variables (return and volatility) equals to number of equations.

We will provide the assessment of the identification for the models from the Table 5.1. The necessary (order) condition for identification is fulfilled for the system of structural equations from Table 5.1, because for each below calculated model the number of excluded independent variables is one and it equals to the number of right-hand side included dependent variable (return or volatility).

The rank condition is $rank(A\phi) = G - 1 = 1$.

$$\mathbf{A}\begin{pmatrix} 1 & volatility_1 & X_{11} & 0 \\ return_2 & 1 & 0 & X_{22} \end{pmatrix}$$

$$\phi_{\mathbf{1}}^{\mathbf{T}} = \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\phi_{\mathbf{2}}^{\mathbf{T}} = \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}$$

Following the calculation from section 3.4 we get:

$$\mathbf{A}\phi_{\mathbf{1}} = \begin{pmatrix} 0 \\ X_{11} \end{pmatrix}$$

$$\mathbf{A}\phi_{\mathbf{2}} = \begin{pmatrix} X_{22} \\ 0 \end{pmatrix}$$

X_{ii} is an explanatory variable.

Thus, the rank condition for the system of equations is satisfied if nor $X_{11} \neq 0$ neither $X_{22} \neq 0$. If these conditions are satisfied, then the system is just-identified. In other words exogenous variables on the right-hand side of the equations have to be different in each equation to fulfill the rank condition.

Let us provide explanatory notes to the Table 5.1. In the first column (Equation 1), there are the explanatory variables that are included into the regression Equation 5.1 and in the second column (Equation 2), the reader might see the explanatory variables from the regression Equation 5.2. In the Table 5.1 there is not provided a comparison for the same explanatory variable

Table 5.1: Comparison of tested models

Equation 1	Equation 2	AIC	BIC	p-value	
				Equation 1	Equation 2
M	Amihud	208 671,8	208 707,0	0,0000	0,0000
M	Amivest	389 042,6	389 077,8	0,0000	0,0000
M	Roll	298 215,3	298 215,3	0,0000	0,0000
M*	**Amihud**	**198 882, 9**	**198 918, 1**	**0, 0000**	**0, 0000**
M*	Amivest	393 041,7	393 076,9	0,0000	0,0000
M*	Roll	294 340,7	294 375,9	0,0000	0,0000
Amihud	M	292 794,6	292 829,8	0,0000	0,0000
Amihud	M*	259 571,8	259 607,0	0,0000	0,0000
Amihud	Amivest	557 436,1	557 471,3	0,0000	0,0000
Amihud	Roll	294 773,5	294 773,5	0,0000	0,0000
Amivest	M	242 001,9	242 037,1	0,0000	0,0000
Amivest	M*	244 261,5	244 296,7	0,0000	0,0000
Amivest	Amihud	266 722,2	266 807,4	0,0000	0,0000
Amivest	Roll	280 392,0	280 427,3	0,0000	0,0000
Roll	M	393 452,4	393 487,6	0,0000	0,0001
Roll	M*	374 064,2	374 099,4	0,0000	0,0007
Roll	Amihud	420 736,0	420 771,3	0,0000	0,2091
Roll	Amivest	676 721,6	676 756,8	0,0000	0,7577

Source: author's computations.

in each equation, because then the order condition would not be satisfied. Also we have not provided a comparison of rounded and exact (”non-rounded”) total number of intraday returns since the 3SLS would then not include any liquidity proxy and the modelling of liquidity is the main purpose of this master's thesis.

M in the Table 5.1 stands for the number of intraday returns within day t. However, as we have standardized the number of intraday returns in the same way as the other variables, the resulting datapoints of M acquire non-integer values. Obviously it does not make a sense, because it is not possible to have non-integer number of intraday returns, thus we have adjusted each datapoint to obtain a rounded number of intraday returns, i.e. M*.

From Table 5.1 we may derive that the best fitted model is with the rounded number of intraday returns in equation one and with Amihud proxy in the equation two. This model has a p-value 0,0000, therefore, its compounded level of confidence is 99,99% (as for most of other models in comparison table, except these with Roll proxy in the first equation) and simultaneously it has the lowest AIC and BIC (198 882,9 and 198 918,1 respectively).

After we have found out the model with the lowest value of AIC and BIC, we have added the variables with lag 1. We have included the both endogenous and exogenous lagged variables and we have made that even for the variables p-value above 0,05 in the initial model. Then we have excluded non-significant lagged variables (with p-value more than 0,05), kept the significant variables and added the variables with lag 2. In the same way we have proceeded on-and-on until we have reached the arbitrary selected maximum lag length $qmax = 5$. Nevertheless, for all examined models, both AIC and BIC were higher than the values of AIC and BIC for the best model from Table 5.1, therefore neither endogenous nor exogenous lagged variables should be included in the model.

We have performed the Augmented Dickey-Fuller test (ADF) to test the null hypothesis of unit-root contained in both endogenous and exogenous variables. We have attached the Table 5.2 with the t-statistic of Augmented Dickey-Fuller test (ADF) for all individual variables.

Table 5.2: ADF test statistic for endogenous and exogenous variables

	RV	return	M	M*	Amihud	Amivest	Roll
lag 1	-102,992*	-158,221*	-112,201*	-114,245*	-112,806*	-128,712*	-154,120*
lag 2	-84,807*	-127,895*	-92,464*	-93,534*	-88,015*	-102,723*	-125,318*
lag 3	-72,467*	-111,343*	-80,685*	-81,305*	-72,979*	-87,849*	-108,867*
lag 4	-65,339*	-101,304*	-71,629*	-72,033*	-63,506*	-76,970*	-96,691*
lag 5	-60,214*	-91,892*	-64,468*	-65,001*	-56,740*	-68,321*	-88,441*

Source: author's computations.

According to Fuller (1976), the critical value is -2,86 for Dickey-Fuller t-distribution for sample size $T > 500$, without any trend and for significance level 5%. Therefore, based on provided data in 5.2 generated by ADF test, we might reject the null hypothesis of unit root for all endogenous and exogenous variables on significance level 5%. The same even applies for significance level 1% ($t = -3,43$). We have reached the same results even when we specified the lag-length and took the trend into the account.

Note that the asterisk by the data in the Table 5.2 signifies the rejection of null hypothesis.

To summarize above described models, we have found out that the best results overall shows up the following 3SLS estimator:

$$r = \alpha_1 RV + \alpha_2 M^* + \epsilon_1, \tag{5.10}$$

$$RV = \beta_1 r + \beta_2 Amihud + \epsilon_2, \tag{5.11}$$

For the final structural Equations 5.10, 5.11, we have predicted the residuals for each equation and performed both tests of unit-root and autocorrelation.

Table 5.3 demonstrates the results of ADF test for residuals of the structural Equations 5.10, 5.11. The asterisk by the data, in the same way as for Table 5.2, stands for the rejection of null hypothesis.

Table 5.3: ADF test for residuals of the final 3SLS model

	residuals from equation 1	residuals from equation 2
lag 1	-158,519*	-158,272*
lag 2	-128,242*	-128,076*
lag 3	-111,810*	-111,601*
lag 4	-101,770*	-101,601*
lag 5	-92,290*	-92,019*

Source: author's computations.

Augmented Dickey-Fuller test for unit-root has the lowest test statistic for lag 5 for both Equations 5.10 and 5.11 and it is 92,290 for Equation 5.10 and -92,019 for Equation 5.11, thus, MacKinnon approximate p-values for Z(t) = 0,0000 for both equations and for all lag-lengths and we may assert that the structural Equations 5.10 and 5.11 do not contain a unit root in their residuals.

Autocorrelation of the disturbances has been tested via Cumby-Huizinga test that is build-in Stata. This test was introduced by Cumby & Huizinga (1992). Cumby-Huizinga test has a null hypothesis that the disturbance is moving average process up to order q and alternative hypothesis that the serial correlation is present at specified lags $> q$. The test automatically displays test statistic for all specified lags and simultaneously also for each lag order.

Cumby-Huizinga test is able to handle the limitations of previously presented autocorrelation tests as for example Breusch-Godfrey test, because it successfully deals with the serial correlation that is even beyond the expected order q. It also takes care of overlapping data that are often present in the financial markets and most importantly, unlike the Breusch-Godrey test, the Cumby-Huizinga test is able to deal with the model that contains endogenous regressors, which is a crucial advantage in our case. Furthermore, Cumby-Huizinga test is even applicable in the situations when the conditional heteroskedasticity in the error process is present.

The Cumby-Huizinga test proved that we do not reject the null hypothesis for lag-length 4 on the significance level 5%, the smallest p-value for first 4 lags is 0,1442, it is a value of the first lag.

Thus, we have suppressed the threat of both strong and weak autocorrelated error terms, i.e. error-term for period t is not correlated with the error term for period s, the autocorrelation does not exist and we may express the disturbances as follows:

$$cov(\mu_t, \mu_s) = E(\mu_t \mu_s) = 0, \tag{5.12}$$

In Table 5.4 we provide the comparison of estimations of the final structural Equations 5.10, 5.11 for OLS, 2SLS and 3SLS.

Table 5.4: OLS, 2SLS and 3SLS estimations

	OLS		**2SLS**		**3SLS**	
			First equation			
	Coeff.	SE	Coeff.	SE	Coeff.	SE
RV	-0,0403***	0,0050	-0,4976**	0,1715	-0,5290***	0,0166
*M**	-0,2560***	0,0048	-0.0215*	0,0086	-0,0197*	0,0082
	$R^2 = 0,0033$		$R^2 = 0,0032$		$R^2 = 0,0032$	
	F=**82,04*** **		F=**54,10*** **		chi^2=**107,67*** **	
			Second equation			
r	-0,0423***	0,0041	-9,0117***	1,0460	-9,1765***	1,0213
Amihud	0,3856***	0,0041	0,1517**	0,0489	0,1415**	0,0469
	$R^2 = 0,1513$		$R^2 = -80,2446$		$R^2 = -83,2258$	
	F=**4402,54*** **		F=**82,56*** **		chi^2=**165,18*** **	

In Table 5.4, we can see that all estimates of coefficients are significant at least at 5% level of significance. The estimates of coefficients for OLS are even all significant on the level of 1 %.

Note that all three models are due to F-test and χ^2 test jointly significant on the confidence level of 99 %. The F-statistic is the test statistic of the overall significance of the individual explanatory variables, thus we may assert that all of the explanatory variables matter.

3SLS achieves lower standard errors than 2SLS. It confirms the theoretical assumption that 3SLS is more efficient than 2SLS. However, overall the OLS estimations have the lowest standard errors.

The interesting fact is that the coefficient of return in the second equation is much higher in the case of 2SLS or 3SLS than with OLS regression. It could be explained by better estimation of relationship between the return and realized variance if we take the endogenity of return into the consideration.

The same might be asserted about the influence of realized variance on the return in the first equation, i.e. the coefficients are again much higher for 2SLS and 3SLS models than for OLS model and the endogenity plays an important role.

We can derive from the comparison of the models that the influence of realized variance on the return is substantially larger than the influence of the rounded number of intraday returns for simultaneous equations. If volatility increases, the return will decrease twice as much. However, completely inverse conclusion can be derived from the OLS model - rounded number of intraday returns has a bigger impact than the volatility.

The similarity in the first equation is in the direction of the influence. Both exogenous variables in the first equation have a negative impact on the return, thus if one exogenous variable rises, the returns falls down.

On the other hand, coefficients of rounded number of intraday returns in the first equation and Amihud Illiquidity in the second equation are higher using the OLS than 2SLS or 3SLS. It signifies that these variables have a bigger influence on the endogenous variables using the OLS model. Specifically rounded number of intraday returns impacts more the return within the single regression than it does within the structural equations. Likewise it is with the Amihud Illiquidity, its effect on the realized variance is larger when it estimates the realized variance only alongside with return.

We can conclude that the realized variance should decrease approximately 9 times as much as the return increases. The return influences the realized variance the most by far. The Amihud Illiqudity influences the realized variance in the case of 2SLS and 3SLS as well. However, the effect is much smaller since the coefficient is only 0,1517 (2SLS) and 0,1416 (3SLS). The volume of the impact is slightly lower than with OLS prediction.

Given the differences in the estimations provided by OLS and simultaneous equations, we have run Hausman's specification test on both equations for OLS and 2SLS.

The Hausman's specification test is based on the test statistics $m = \hat{q}^T[var(\hat{q})]^{-1}\hat{q}$ with asymptotical χ^2_k distribution where k is the number of right-hand side endogenous variables.

Using Gretl econometric program, for the first equation we get asymptotic test statistic

$\chi_1^2 = 0,329113$ with p-value $0,566182$. We therefore cannot reject the null hypothesis $H_0 = E(Z_i^T u_i) = 0$ and endogenous variable realized variance is probably correlated with the error tem. Thus OLS should be used as it shows the dependence of the return on the realized variance and the number of intraday returns better than it does the simultaneous equations. However, for the second equation we get the test statistic $\chi_1^2 = 8223,67$ with p-value 0 and thus, we reject the null hypothesis that OLS estimates are consistent. 2SLS or 3SLS estimators can be used and they should be consistent in the estimation of the dependence of the realized variance on the return and Amihud Illiquidity.

Nevertheless, the key conclusion that is derived from the model is that our hypotheses is not confirmed. The liquidity truly enters into the return-volatility relationship and influences these variables, even though it is rather marginal impact - the magic triangle interacts. However, against general perception and our hypotheses, the model shows up that the lower liquidity does not signify the higher realized variance. On the contrary, lower liquidity signifies lower realized risk and, through proved risk-return relationship in the structural equations, also lower return. This conclusion has been suggested by all three models.

5.2 The lower liquidity signifies higher frequency of jumps

We have run the VBA code through the dataset and obtained daily realized variance and daily bi-power variation, both calculated on the basis of high-frequency data. The equations, we have used for VBA calculation, were already presented in the subsections 3.2 and 3.3.

5.2.1 Interaction of realized variance and bi-power variation

We have shown in subsection 3.3 that daily BV derived from high-frequency data converges to the integrated variance unaffected by jumps, it implies that the result of $RV_t - BV_t$, where t means a day, denotes a volume of the jump component in particular daily realized variance. Barndorff-Nielsen & Shephard (2004) and Huang & Tauchen (2005) suggested following jump statistic to reveal a jump on each day t.

$$z_t = \frac{\frac{RV_t - BV_t}{RV_t}}{\sqrt{(\frac{\pi^2}{4} + \pi - 5)\frac{1}{M}\frac{TP_t}{BV_t^2}}}, \tag{5.13}$$

$$TP_t = M\mu_{4/3}^{-3}\frac{M}{M-2}\sum_{i=3}^{M}|r_{t,i-2}|^{4/3}|r_{t,i-1}|^{4/3}|r_{t,i}|^{4/3}, \tag{5.14}$$

where $\mu_k = 2^{k/2}\Gamma[(k+1)/2]/\Gamma(1/2)$ is a normalizing term and $\Gamma(p) = \int_0^\infty t^{p-1}e^{-t}dt$ for any positive p. Since

$$z_t \overset{M\to\infty}{\longrightarrow} N(0,1)\ , \tag{5.15}$$

when we label the continuous and jump volatility components as C_t and J_t respectively, we will calculate both C_t and J_t in the following way:

$$C_t = 1_{z_t \leq z_\alpha} RV_t +, 1_{z_t > z_\alpha} BV_t, \tag{5.16}$$

$$J_t = 1_{z_t > z_\alpha}(RV_t - BV_t), \tag{5.17}$$

where α is the upper $99,99\%$-quantile of N(0,1).

Cheng *et al.* (2013) suggest upper 99,99% quantile, however, in our model we will employ more common 95% quantile.

Let us just remind that estimated jump in the Equation 5.17, cannot be negative, because RV_t is supposed to be higher than BV_t, in other case estimated jump should be zero. However, in our model we just take into the account whether the jump occurred rather than the precise volume of the jump, thus we do not have to consider this potential issue.

In this moment based on daily RV and BV we have a set of dummy variables, which denote whether the jump occurred in a particular day. In the following Table 5.5 we have provided the summary of number of jumps for a particular share. Since our data are strongly balanced we have 504 observations for each of 98 shares. To see the name of the company behind the share's symbol, refer to Appendix A.

We might see from Table 5.5 that the least within day jumps from 98 shares in the examining period occurred by Facebook shares (82 jumps in 504 active trading days) and the most within day jumps by Lockheed-Martin (190). According to our hypothesis it would mean that for these shares, the liquidity is the highest and lowest respectively. Median is 152 jumps. In total there is 14 649 day jumps within 49 392 data points, it tells us that the jump occurs in average every third or fourth day (49 392/14649 = 3,372).

We will try to explain this set of dummy variables due to liquidity, volatility, returns and lags of all these three variables. Since our dependent variable is binary (dummy variable of daily jumps takes value 0 or 1), we will use nonlinear regression models for panel data - logit and probit.

5.2.2 Logit & Probit

Logit model estimates the probability of dummy variable to be 1, i.e. the probability that a jump in particular day takes place.

Using a latent variable framework, we might use the panel binary choice model as:

$$y^*_{it} = x_{it}\beta + c_i + u_{it}, \tag{5.18}$$

$$y_{it} = 1[y^*_{it} > 0], \tag{5.19}$$

and

$$Pr(y_{it} = 1|x_{it}, c_i) = G(x_{it}\beta + c_i), \tag{5.20}$$

Table 5.5: Number of jumps per share during the 504 active market days since 1/1/2013 until 31/12/2014

symbol	# jumps	*symbol*	# jumps	*symbol*	# jumps	*symbol*	# jumps
aapl	110	cvs	166	intc	130	pg	147
abbv	179	cvx	155	jnj	126	pm	157
abt	182	dd	160	jpm	116	qcom	160
acn	158	dis	138	ko	172	rtn	158
aig	129	dow	154	lly	180	sbux	120
all	160	dvn	133	lmt	190	slb	116
amgn	145	ebay	134	low	152	so	164
amzn	109	emc	144	ma	129	spg	144
apa	155	emr	177	mcd	127	t	149
apc	127	exc	165	mdlz	180	tgt	154
axp	174	f	152	mdt	170	twx	180
ba	137	fb	82	met	152	txn	159
bac	156	fcx	126	mmm	184	unh	163
bax	172	fdx	153	mo	140	unp	178
biib	127	foxa	175	mon	173	ups	151
bk	158	gd	145	mrk	172	usb	161
bmy	179	ge	131	ms	128	utx	176
c	105	gild	115	msft	121	v	161
cat	132	gm	140	nke	147	vz	141
cl	175	gs	143	nov	162	wba	145
cmcsa	181	hal	114	nsc	161	wfc	139
cof	157	hd	147	orcl	162	wmt	150
cop	133	hon	167	oxy	161	xom	116
cost	171	hpq	126	pep	158		
csco	143	ibm	141	pfe	130		

Source: author's computations.

where $G(.)$ is either the standard normal cumulative distributive function (probit) or the logistic cumulative distributive function (logit).

Logit and probit models are calculated both on the same basis, therefore naturally they provide similar results. The difference is in their distribution. Logit has a cumulative standard logistic distribution (F), meanwhile probit has a cumulative standard normal distribution (Φ). Thus using the logarithm values, that was described in the first part of section 5, is especially important for probit, as the logit already has logistic distribution included.

5.2.3 Fixed effects & Random effects model

As we have a panel data, we have to decide which model will better suit to us, fixed effects (FE) or random effects (RE) model. FE model analyses the connection between exogenous and endogenous variable within a subgroup, i.e. within one share in our case. Every share has its own attributes that might or might not impact the possibility of within day jump occurrence. With FE model, we assume the correlation between subgroup's disturbance and endogenous variable and take it into the consideration. Note that fixed effects model is without a constant since everything that is constant within a panel unit is necessarily eliminated from the regression model.

FE model is able to eliminate effects of time-invariant attributes so we might appraise the real effect of exogenous variable on the endogenous variable. Notice please that these time-invariant attributes are unique to the subgroup and they are not supposed to be correlated with other subgroups' attributes.

In the case the disturbances are correlated with other subgroups' disturbances, FE model is not suitable and we should employ RE instead to get more realistic conclusions. RE model assumes that differences across subgroups have an impact on the endogenous variable. To make a decision which model to use we will run Hausman test that will be described later.

Initially, we have assumed that our model should be with fixed effects, because each share might include its own attributes. Our assumption is supported by Allison (2009), the author states two conditions for using fixed effects in logit models.

- Endogenous variable has to occur at least two times for each subgroup.
- Exogenous variable has to vary across the time for considerable part of the subgroups.

Our dataset complies both conditions. As we have seen in the Table 5.5, within day jump occurs at least twice by each share and as we went through the dataset, we might say that exogenous variables vary substantially for each share.

Allison (2009) points out that fixed effects model is only concerned with within-subgroup of differences and eliminates any differences between individual subgroups. It might cause a problem, because if variables vary vastly between individuals and vary rarely for each subgroup, then fixed effects model is inaccurate and has a great standard errors. However, given our dataset, we assume that between-subgroup variation is not a significantly higher than within-subgroup variation, therefore standard errors of the coefficients are not too large to accept. Our assumption is based on the fact that our dataset is consisted of highly traded shares on the biggest stock markets on the world. Besides, Allison (2009) still prefers fixed effects model over random one since it is better protected against omitted variable bias.

We have used Hausman test to confirm our assumption. Essentially it tests whether the unique errors (ϵ_i) are correlated with regressors. The null hypothesis that difference in coefficients is not systematic has been rejected on 99 % level of confidence for random effects model. Given Torres-Reyna (2011) and "persuasive" Hausman test we have a very strong argument to employ fixed effects instead of random effects. Despite the fact, we will include some results from the models with random effects to obtain a more general comparison.

By using the fixed effects model we have to pay attention to two shortcomings that nonlinear fixed effects models have.

The first shortcoming, the practical one, is connected with the estimation of thousands of dummy variables. However, Greene *et al.* (2002) asserts that the estimation by fixed effects model is feasible in panel data with great number of subgroups. Given our sample of 98 shares, we perceive this shortcoming as treated. The second shortcoming, the methodological one, seems to cause more serious problems as it concerns statistical properties of the estimator, Greene *et al.* (2002) claim that probit gets inconsistent with fixed effects, probit is able to model only random effects (both with and without a constant). It is even not possible to run a fixed effects model for probit in Stata.

5.2.4 Stationarity

Further we have employed unit root tests for panel data. We have performed commonly used Fisher-type unit-root test based on Augmented Dickey-Fuller test for all endogenous and exogenous variables. Besides ordinary Fisher type, we have also carried out an adjusted testing that included time trend as we tended to embrace the historical development of shares on the stock markets. We have used step by step lags 1 to lags 5. Interestingly, all tests rejected the null hypothesis that panels contain unit roots on a confidence level 99 %. However, Fisher-type tests are better for the estimation of unbalanced panels and we have already stated earlier in Chapter 5 that our data are strongly balanced. Besides, we know from the literature that volatility has a long memory, therefore we have decided to perform other tests to obtain a clearer perspective on the issue of unit root.

Classical statistical methods are designed to reject the null hypothesis only when the evidence against null is sufficiently overwhelming. However, unit-root tests are usually not very powerful against alternative hypothesis. We have employed Hadri LM test (Hadri 2000), which has, unlike the Fisher-type, the null hypothesis that the process is stationary. Moreover Hadri LM test is designed for balanced data and since we have those, it should be more suitable.

Table 5.6: Hadri unit-root test

Variable	Trend excluded		Trend included	
	z-statistic	p-value	z-statistic	p-value
Jump	2,02*	0,0217	0,90	0,1842
Return	-1,48	0,9304	0,72	0,2348
Realized variance	126,63*	0,0000	110,50*	0,0000
Amihud	381,88*	0,0000	89,14*	0,0000
Amivest	167,45*	0,0000	95,45*	0,0000
Roll	7,60*	0,0000	6,65*	0,0000

Source: author's computations.

As expected, Table 5.6 shows that Hadri LM test for realized variance and liquidity proxies rejected null hypothesis of stationarity on the confidence level of 99 %. However, p-value 0,9304 (0,2348 for model with trend involved) for return suggests that daily returns based on high frequency data are truly stationary, it means that previous days' returns do not influence the fact whether the jump occurs within the current day. For within-day jump null has been rejected on 95% confidence level, but when we have included trend, we have obtained p-value 0,1842, thus the stationarity is not rejected. Once more the asterisk by the data signifies the rejection of null hypothesis.

As the results from unit-root tests were ambiguous, we have decided to test the model, except the return, in both ways, i.e. with and without lags.

5.2.5 Approaches to modelling

To obtain the best model regarding our dataset, we have modeled various regressions employing logit, probit, fixed effects model, random effects model (with and without a constant) and GLS. We have tested the models with and without lags of endogenous and exogenous variables. We have compared the results and below introduced a table with a comprehensive comparison.

Since substantial number of models proved to be valid, i.e. p-values of all variables were below 0,5 and models were statistically jointly significant on the level of confidence 99 %, we had to come up with additional method to appraise the particular model. To assess individual models we have again used two criterions provided by Stata - AIC and BIC.

Chen & Tsurumi (2010) provided a comparison of criterions to appraise the suitability of individual probit and logit models. Authors claimed that if data are balanced, none of the criterions are able to distinguish the probit and logit models. Generally the distribution of binary data is unknown, Chen & Tsurumi (2010) declare that the financial return data are leptokurtic.

During the modeling we have proceeded in the following way. Firstly, we have tested models considering individual combinations of the examined exogenous variables (return, realized variance and three liquidity proxies). Subsequently, for the best combination according to criterions we have included first lag for all exogenous and endogenous variables. We have found out the significant variables and for the model with the lowest AIC and BIC we have added second lag for the variables that had already first lag included in the equation. We have continued in the same way up to 5 lags. We have performed this procedure for logit and probit models.

We involved return, RV and at least one liquidity proxy in each combination as we tended to embrace the magical triangle of investments. During the testing it has been confirmed that the models that included all three "edges of magical triangle" proved significantly better results.

We have also tested both individual liquidity proxy and set of liquidity proxies in the model as we were concerned whether more liquidity proxies might cause a problem. Eventually set of liquidity proxies proved to show up better results, which means that together the liquidity proxies are better estimator of within-day jump than if separated.

F-test with null hypothesis that all of the fixed effects' intercepts are zero has been employed. Again, the null has been rejected on the level of confidence 99 %.

Finally, we reached to the point, where the best models according to criterions proved to be the following ones. The best logit model generated from the above described procedure is:

$$y_{it} = F(x_{it}^T\alpha) + u_{it}, \tag{5.21}$$

$$F(x_{it}^T\alpha) = \frac{1}{1 + exp\{x_{it}^T\alpha\}}, \tag{5.22}$$

$$\begin{aligned} x_{it}^T\alpha + u_{it} &= \alpha_0 + \alpha_{1t}r_{it} + \alpha_{2t}RV_{it} + \alpha_{3t}RV_{it-1} + \alpha_{4t}RV_{it-2} + \alpha_{5t}RV_{it-3} + \\ &+ \alpha_{6t}RV_{it-4} + \alpha_{7t}RV_{it-5} + \alpha_{8t}Amivest_{it} + \alpha_{9t}Roll_{it} + u_{it}, \quad (5.23) \end{aligned}$$

and the best probit model generated by the procedure is:

$$y_{it} = F(x_{it}^T\beta) + u_{it}, \tag{5.24}$$

$$F(x_{it}^T\beta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_{it}^T} exp\left\{\frac{t^2}{2\sigma^2}\right\} \mathrm{d}t, \tag{5.25}$$

$$\begin{aligned} x_{it}^T\beta + v_{it} &= \beta_0 + \beta_{1t}r_{it} + \beta_{2t}RV_{it} + \beta_{3t}RV_{it-1} + \beta_{4t}RV_{it-2} + \beta_{5t}RV_{it-3} + \\ &+ \beta_{6t}RV_{it-4} + \beta_{7t}RV_{it-5} + \beta_{8t}Amivest_{it} + \beta_{9t}Amivest_{it-3} + \\ &+ \beta_{10t}Roll_{it} + v_{it}, \quad (5.26) \end{aligned}$$

where $i = 1, \ldots, 98$ is the number of cross-sectional units, i.e. shares, $t = 1, \ldots, 504$ is the time-series dimension, i.e. period and σ is the variance of the associated normal distribution.

We might see that the return, realized variance and Amivest and Roll liquidity proxies are the best estimators of the occurrence of a within day jump. Furthermore, we might assert that the volatility expressed by realized variance has truly a "long memory". Meanwhile the other lagged variables (except Amivest) do not significantly influence the occurrence of the jump.

In the Tables 5.7 and 5.8 we have provided a comparison of models based on Equations 5.23 and 5.26. The models regarded fixed effects and random effects model, presence and absence of the constant in the equation and GLS estimator. Table 5.7 shows the comparison of models based on Equation 5.23. Table 5.8 demonstrates models based on Equation 5.26. The figures in the Tables 5.7 and 5.8 are rounded to one decimal place.

Table 5.7: Comparison of tested models #1

Model		AIC	BIC
logit	FE	57 894,4	**57 973, 7**
	RE with const	58 840,7	58 937,6
	RE no const	59 019,9	59 108,0
probit	FE	N/A	N/A
	RE with const	58 849,2	58 946,0
	RE no const	59 033,9	59 121,9
GLS	FE	61 197,9	61 285,9

Source: author's computations.

Table 5.8: Comparison of tested models #2

Model		AIC	BIC
logit	FE	**57 886, 5**	57 974,6
	RE with const	58 822,7	58 928,4
	RE no const	58 938,5	59 035,4
probit	FE	N/A	N/A
	RE with const	58 831,0	58 936,7
	RE no const	58 950,6	59 047,5
GLS	FE	61 189,5	61 286,4

Source: author's computations.

Notice that in the random effects GLS regression there is no likelihood information and therefore, it is not possible to assess the model due to Akaike or Bayesian information criterion. Furthermore, there is no nonconstant option for the random effects GLS regression in Stata. Altogether it implies that GLS regression is carried out only for fixed effects model. Furthermore, as stated in the subsection 5.2.3, probit gets inconsistent with FE and thus it is not possible to model this estimation in Stata.

We might see from the Tables 5.7 and 5.8 that conditional fixed effects logistic regression' results dominate the criterions, Equation 5.23 has the lowest BIC and Equation 5.26 has the lowest AIC in the FE logit model.

Following Acquah (2010) we have chosen BIC to be a crucial parameter for an overall assessment of the models. Therefore, the resulting model of this master thesis is fixed effects logit model with Equation 5.23.

Acquah (2010) carried out a testing involving Monte Carlo method that proved that BIC consistently outperforms AIC in the selection of appropriate asymmetric price relationships within enormous datasets. The author states that AIC prevails BIC only in marginal cases, when the sample is small and large noise levels are present. We might assert that our dataset is large since it is composed of 49 392 datapoints and large noise levels are generally not present on the most developed stock-exchanges in the non-crisis times.

AIC tends to be too liberal and favours more complex, wrong models over a simpler, true models. AIC is trying to discover the best estimation model for unknown data generating process and therefore, it does not converge in probability to the true model, meanwhile BIC does converge as N goes to infinity.

5.2.6 Final model description

Appendix B captures the output calculated in STATA for our final equation. The note "multiple positive outcomes within groups encountered" signifies that there are more than one positive outcomes in the dataset, it might be a problem for some analyses, however, not for our case, because within day jumps might (and they are) present several times during the investigated period.

Logistic regression uses maximum likelihood estimator. Maximum likelihood is an iterative procedure, thus log likelihoods for each iteration are listed further. First iteration is the logistic likelihood of null hypothesis, i.e. log likelihood of the model without any exogenous variables. For the next iterations exogenous variables are included into the model in the way to maximize the logistic likelihood. When the difference between iterations is very low, then we might say that the model "converged", iterations stop and the outcomes are provided. The value of log likelihood in the last iteration is the one that is used in the model. The particular value has no significant meaning in the calculation, it might be rather used to compare tested models. However, as stated in subsection 5.2.5 we have compared the models according to AIC and BIC instead.

The number of observations is 49 392 as mentioned in Data section. There is 504 observations in every group, thus we may see that the data are strongly balanced. Group variable that is used for fixed effects model is share. Remember that the values of variable "share", i.e. the names of the share on the stock market, do not vary within the subgroup, thus this variable is withdrawn from the equation, because the constant is not able to explain the variability in the endogenous variable.

LR chi2 (9) is the likelihood ratio (LR) of chi-square test. It can be calculated as 2*(Iteration 0 - Iteration 3), i.e. it shows us the value of the difference between first and last log likelihood multiplied by two. In the parenthesis there is a number of degrees of freedom. We have used 9 exogenous variables, thus we have 9 degrees of freedom.

$Prob > chi2$ shows the probability of the null hypothesis, i.e. the probability that model with no-effect exogenous variables on the endogenous variable may reach the chi-square statistic (1 112,51). Actually this is the p-value and thus we can say that the model is statistically significant on 99,99 % confidence level as the p-value is less than 0,0000. It signifies that the model is sufficiently comprehensive to capture the factors that influence within day jumps. We had said that the model is significant even when the p-value would be less than 0,05.

Note that there is no R-squared mentioned as logistic regression does not have an equivalent to R^2 as we know that for example by OLS, even though there was an effort to come up with some.

We may see that two-tail p-value is lower than expected significance level 0,05 for all variables in the model, respectively z-statistic is larger than the expected benchmark 1,65, i.e. the value of 95 percentile point of the normal distribution. This fact is not surprising as it was one of our conditions when we have chosen the appropriate model. All non-lagged exogenous variables (return, realized variance, Amivest, and Roll) are statistically significant on the confidence level of 99,99 %.

Let us remind that the estimated coefficients in logit regression are log-odds units, it means that it is not possible to read them in the way it is common with OLS. As the log-odds units are rather difficult to interpret, we have used another command in Stata to generate the odds-ratios (OR), the results are provided in Appendix B, as well.

The odds-ratios represent the odds of jump-occurrence when exogenous variable increases about 1 unit. When $OR > 1$ the odds of jump-occurrence rises, on the other hand, when $OR < 1$ the odds of jump-occurrence declines. For return, realized variance and Amivest the ORs are greater than 1 (coefficients in the Table B.1 are positive), it means that the higher the value of these variables is, the higher the probability of within-day jump occurrence is. In contrary to lags of realized variance and Roll liquidity proxy, which ORs are less than 1 (coefficients in the Table B.1 are negative), i.e. with increasing value of these exogenous variables, within-day jump occurrence declines. Therefore our hypothesis that the lower liquidity signifies higher frequency of jumps is ambiguous, it is confirmed only partially - due to Roll. However, Amivest liquidity proxy stands against the hypothesis.

Most significant exogenous variable that influences the within-day jump is realized variance, it has the biggest z statistic (30,91), largest confidential interval and its OR is 1,563, thus the positive/negative change in the RV affects the rising/declining probability of jump occurrence the most (from the presented predictors).

However, on the basis of our results we might claim that all the exogenous variables in the model are significant predictors of the future development of within-day jumps.

Chapter 6

Conclusion

This master thesis concentrates on the influence of liquidity measures on the prediction of volatility and given the magic triangle phenomena subsequently on the expected return. In other words, we investigate if it is of any use to include the liquidity into the investors' decision-making process about the potential investments.

We assumed that deviations from standard return-risk relation are caused by liquidity, lower liquidity suggests higher realized risk and return and moreover that lower liquidity signifies higher frequency of jumps.

Liquidity measures have been chosen according to Goyenko *et al.* (2009), these liquidity measures are price impact proxies Amihud Illiquidity, Amivest Liquidity and Roll. Goyenko *et al.* (2009) proposed the low-frequency versions of these proxies. We have had come up with the high-frequency's adjustments as we have obtained the high frequency data with 5-min sampling frequency.

Dataset used for the modeling was consisting of 98 shares that were traded on S&P 100 as of 15th May 2015. The time range was from 1st January 2013 to 31st December 2014, in total the dataset had 49 932 data-points.

We have used realized variance instead of unobservable quadratic variation as volatility measure. As quadratic variation consists of a term representing the continuous price path and a term representing the within-day jumps and the realized variance converges in probability to quadratic variation (Andersen *et al.* 2000), we were able to separate the jump from quadratic variation due to bi-power variation.

In the empirical section, the 3SLS model for hypothesis regarding the relation among return, risk and liquidity has been used. The best resulting model has been compared with 2SLS and OLS regression. We have found out that the liquidity truly enters into the return-volatility relationship and influences these variables - the magic triangle interacts. However, contrary to our hypothesis, the model shows up that lower liquidity signifies lower realized risk. This inference has been suggested by all 3SLS, 2SLS and OLS models.

For the jump-occurrence modeling, we have utilized various regressions regarding logit, probit, fixed effects model, random effects model (with and without a constant) and GLS. We have tested the models with and without lags of endogenous and exogenous variables, provided a comprehensive comparison and extensively commented on the final model that prevailed the other ones.

Our hypothesis concerning the jumps was confirmed only for one of two liquidity proxies in the final model, for Roll specifically. As the most significant variable influencing the within-day jump proved to be the realized variance.

The examined hypotheses that are widely accepted by economists might be confirmed in the future by larger dataset that would cover more than two years. Moreover, some development markets could be also concerned.

Bibliography

Acquah, H. (2010): "Comparison of akaike information criterion (aic) and bayesian information criterion (bic) in selection of an asymmetric price relationship." *Journal of Development and Agricultural Economics* **2(1)**: pp. 001–006.

Allen, K., K. Daniels, D. Kopp, & B. Murdock (2004): "Analysis of 2004 Political Futures Markets." Mimeo.

Allison, P. D. (2009): *Fixed effects regression models*, volume 160. SAGE publications.

Amihud, Y. (2002): "Illiquidity and stock returns: cross-section and timeseries effects." *Journal of Financial Markets* **5**: pp. 31–56.

Andersen, T. G., L. Benzoni, & J. Lund (2002): "An empirical investigation of continuous-time equity return models." *The Journal of Finance* **57(3)**: pp. 1239–1284.

Andersen, T. G., T. Bollerslev, F. X. Diebold, & P. Labys (2000): "Exchange rate returns standardized by realized volatility are (nearly) gaussian." *Technical report*, National bureau of economic research.

Barberis, N., A. Shleifer, & R. Vishny (1998): "A model of investor sentiment." *Journal of Financial Economics* **49**: pp. 307–343.

Barndorff-Nielsen, O. E. & N. Shephard (2004): "Power and bi-power variation with stochastic volatility and jumps." *Journal of Financial Econometrics* **2**: pp. 1–37.

Barunik, J. (2013a): "Quantitative Finance I: Lecture 7 (High-frequency financial models - Microstructure noise)." `http://staff.utia.cas.cz/barunik/files/QFI/QF_I_Lecture7.pdf`. 2014-06-25.

BARUNIK, J. (2013b): "Quantitative Finance I: Lecture 9 (High-frequency financial models II - Introduction to Realized Measures)." `http://staff.utia.cas.cz/barunik/files/QFI/QF_I_Lecture9.pdf`. 2014-06-25.

BATES, D. S. (2000): "Post-'87 crash fears in the s&p 500 futures option market." *Journal of Econometrics* **94(1)**: pp. 181–238.

BLACK BOOK (1996): *Black Book - The Future of Money Management in America (1997 Edition)* p. 22.

BUNGE, J. & J. STRASBURG (2012): "Social Network's Debut on Nasdaq Disrupted by Technical Glitches, Trader Confusion." *The Wall Street Journal* **49**: pp. 307–343.

Caltech (2015) (2015): "Caltech Market Data Guide." `http://quant.caltech.edu/historical-stock-data.html`. 2015-07-28.

CHEN, G. & H. TSURUMI (2010): "Probit and logit model selection." *Communications in Statistics-Theory and Methods* **40(1)**: pp. 159–175.

CHENG, A., K. DAS, & T. SHIMATANI (2013): "Central bank intervention and exchange rate volatility: Evidence from japan using realized volatility."

CHO, D. D. & E. KUVVET (2015): "Dollar-cost averaging: The trade-off between risk and return." *Journal of Financial Planning* **28(10)**: pp. 52 – 58.

CHVÁTALOVÁ, Z., J. HŘEBÍČEK *et al.* (2013): "Computational finance and finance economics with maple." *International Journal of Mathematical Models and Methods in Applied Sciences* **7**: pp. 541–550.

COOPER, S. K., J. C. GROTH, & W. E. AVERA (1985): "Liquidity, exchange listing, and common stock performance." *Journal of Economics and Business* **37**: pp. 319–33.

CUMBY, R. E. & J. HUIZINGA (1992): "Investigating the correlation of unobserved expectations: Expected returns in equity and foreign exchange markets and other examples." *Journal of Monetary Economics* **30(2)**: pp. 217–253.

DANIEL, K., D. HIRSHLEIFER, & A. SUBRAHMANYAM (1997): "A theory of overconfidence, self-attribution, and security market under- and overreactions." Mimeo.

DeBondt, W. & R. Thaler (1985): "Does the stock market overreact?" *Journal of Finance* **40**: pp. 793–805.

Eraker, B., M. Johannes, & N. Polson (2003): "The impact of jumps in volatility and returns." *The Journal of Finance* **58(3)**: pp. 1269–1300.

Erbar, P. (2014): "20 Largest Stock Exchanges in the World." `http://www.insidermonkey.com/blog/trading-places-the-20-largest-stock-exchanges-in-the-world-335310/`. 2015-07-28.

Evans, T. (2006): "Efficiency tests of the UK financial futures markets and the impact of electronic trading systems." *Applied Financial Economics* **16**: pp. 1273–1283.

Fama, E. F. (1998): "Market efficiency, long-term returns, and behavioral finance." *Journal Of Financial Economics* **49**: pp. 283–306.

Figlewski, S. & X. Wang (2000): "Is the "leverage effect" a leverage effect?" NYU Stern School of Business.

Fisher, I. N. & G. R. Hall (1969): "Risk and corporate rates of return." *The Quarterly Journal of Economics* pp. 79–92.

Fuller, W. A. (1976): "Introduction to statistical time series, new york: Johnwiley." *FullerIntroduction to Statistical Time Series1976* .

Goyenko, R. Y., C. W. Holden, & C. A. Trzcinka (2009): "Do liquidity measures measure liquidity?" *Journal of Financial Economics* **92**: pp. 153–181.

Greene, W., C. Han, & P. Schmidt (2002): "The bias of the fixed effects estimator in nonlinear models." *Unpublished Manuscript, Stern School of Business, NYU* .

Hadri, K. (2000): "Testing for stationarity in heterogeneous panel data." *The Econometrics Journal* pp. 148–161.

Han, Y. & D. Lesmond (2011): "Liquidity biases and the pricing of cross-sectional idiosyncratic volatility." *Review of Financial Studies* **24(5)**: pp. 1590–1629.

HASBROUCK, J. (2009): "Trading costs and returns for US equities: estimating effective costs from daily data." *Journal of Finance* **64**: pp. 1445–1477.

HIMMELMANN, A., D. SCHIERECK, M. SIMPSON, & M. ZSCHOCHE (2012): "Long-term reactions to large stock price declines and increases in the European stock market: a note on market efficiency." *Journal Of Economics & Finance* **36**: pp. 400–423.

HITCHNER, J. R. (2010): *Financial Valuation,+ Website: Applications and Models*, volume 545. John Wiley & Sons.

HODNETT, K. & H. HENG-HSING (2012): "Capital Market Theories: Market Efficiency Versus Investor Prospects." *International Business & Economics Research Journal* **11**: pp. 849–862.

HUANG, X. & G. TAUCHEN (2005): "The relative contribution of jumps to total price variance." *Journal of financial econometrics* **3(4)**: pp. 456–499.

KOOP, G. (2014): "Chapter 8: Regression with Lagged Explanatory Variables." `http://personal.strath.ac.uk/gary.koop/Oheads_Chapter8.pdf`. 2015-07-28.

LIPTAK, S. (2012): *Forecasting realized volatility: Do jumps in prices matter?* Master's thesis, Charles University in Prague, the Czech Republic.

MANCINI, L., A. RANALDO, & J. WRAMPELMEYER (2013): "Liquidity in the foreign exchange market: Measurement, commonality, and risk premiums." *Journal of Finance* **68(5)**: pp. 1805 – 1841.

MYKLAND, P., N. SHEPHARD, & K. SHEPPARD (2010): "Econometric analysis of financial jumps using efficient bipower variation."

NEUMANN, M., I. BÖBEL, & A. HAID (1979): "Profitability, risk and market structure in west german industries." *The Journal of Industrial Economics* **27(3)**: pp. 227–242.

NOVAK, J. (2012): "Financial Markets: Lecture 2 (Efficient Markets)." Mimeo.

OBERHOFER, W. & J. KMENTA (1973): "Estimation of standard errors of the characteristic roots of a dynamic econometric model." *Econometrica: Journal of the Econometric Society* pp. 171–177.

PASTOR, L. & R. STAMBAUGH (2003): "Liquidity risk and expected stock returns." *Journal of Political Economy* **111**: pp. 642–685.

POOTER, M. d., M. MARTENS, & D. v. DIJK (2008): "Predicting the daily covariance matrix for s&p 100 stocks using intraday data-but which frequency to use?" *Econometric Reviews* **27(1-3)**: pp. 199–229.

PROKOPOVA, D. (2003): *Moznosti zhodnoceni a investovani rizikove averzniho klienta*. Master's thesis, Bankovni institut vysoka skola Praha, the Czech Republic.

RHOADS, R. (2014): *Trading Weekly Options: Pricing Characteristics and Short-term Trading Strategies*. John Wiley & Sons.

ROBERT MORRIS ASSOCIATES (1988): *A Guide to Analyzing Foreign Banks*. Robert Morris Associates.

ROLL, R. (1984): "A simple implicit measure of the effective bid-ask spread in an efficient market." *The Journal of Finance* **39**: pp. 1127–1139.

ROSENBERG, A. (2013): "Share prices are soaring, but splits aren't coming back." `http://www.cnbc.com/2013/12/12/share-prices-are-soaring-but-splits-arent-coming-back.html`. 2015-06-25.

RYU, A. (2011): "Beta Estimation Using High Frequency Data." .

SRIBNEY, W., V. WIGGINS, & D. DRUKKE (2015): "Negative and missing R-squared for 2SLS/IV."

TORRES-REYNA, O. (2011): "Panel data analysis fixed & random effects." *Princeton University, Data and Statistical Services presentation* .

VESELA, J. (2011): *Investovani na kapitalovych trzich*. Prague: Wolters Kluwer.

XUE, L. (2014): *Liquidity-adjusted Expected Shortfall*. Ph.D. thesis, New York University.

ZELLNER, A. & H. THEIL (1962): "Three-stage least squares: simultaneous estimation of simultaneous equations." *Econometrica: Journal of the Econometric Society* pp. 54–78.

Appendix A

List of S&P100 shares

Table A.1: List of S&P100 shares as at 15th May 2015

symbol	company
aapl	Apple Inc.
abbv	AbbVie Inc.
abt	Abbott Laboratories
acn	Accenture plc
aig	American International Group Inc.
all	Allstate Corp.
amgn	Amgen Inc.
amzn	Amazon.com
apa	Apache Corp.
apc	Anadarko Petroleum Corporation
axp	American Express Inc.
ba	Boeing Co.
bac	Bank of America Corp
bax	Baxter International Inc
biib	Biogen Idec
bk	Bank of New York
bmy	Bristol-Myers Squibb
c	Citigroup Inc
cat	Caterpillar Inc
cl	Colgate-Palmolive Co.
cmcsa	Comcast Corporation
cof	Capital One Financial Corp.
cop	ConocoPhillips
cost	Costco
csco	Cisco Systems
cvs	CVS Caremark

symbol	company
cvx	Chevron
dd	DuPont
dis	The Walt Disney Company
dow	Dow Chemical
dvn	Devon Energy
ebay	eBay Inc.
emc	EMC Corporation
emr	Emerson Electric Co.
exc	Exelon
f	Ford Motor
fb	Facebook
fcx	Freeport-McMoran
fdx	FedEx
foxa	Twenty-First Century Fox, Inc
gd	General Dynamics
ge	General Electric Co.
gild	Gilead Sciences
gm	General Motors
goog	Google Inc.
googl	Google Inc.
gs	Goldman Sachs
hal	Halliburton
hd	Home Depot
hon	Honeywell
hpq	Hewlett Packard Co
ibm	International Business Machines
intc	Intel Corporation
jnj	Johnson & Johnson Inc
jpm	JP Morgan Chase & Co
ko	The Coca-Cola Company
lly	Eli Lilly and Company
lmt	Lockheed-Martin
low	Lowe's
ma	Mastercard Inc
mcd	McDonald's Corp
mdlz	Mondelez International
mdt	Medtronic Inc.
met	Metlife Inc.
mmm	3M Company
mo	Altria Group
mon	Monsanto
mrk	Merck & Co.

symbol	company
ms	Morgan Stanley
msft	Microsoft
nke	Nike
nov	National Oilwell Varco
nsc	Norfolk Southern Corp
orcl	Oracle Corporation
oxy	Occidental Petroleum Corp.
pep	Pepsico Inc.
pfe	Pfizer Inc
pg	Procter & Gamble Co
pm	Phillip Morris International
qcom	Qualcomm Inc.
rtn	Raytheon Co (NEW)
sbux	Starbucks Corporation
slb	Schlumberger
so	Southern Company
spg	Simon Property Group, Inc.
t	AT&T Inc
tgt	Target Corp.
twx	Time Warner Inc.
txn	Texas Instruments
unh	UnitedHealth Group Inc.
unp	Union Pacific Corp.
ups	United Parcel Service Inc
usb	US Bancorp
utx	United Technologies Corp
v	Visa Inc.
vz	Verizon Communications Inc
wba	Walgreens Boots Alliance
wfc	Wells Fargo
wmt	Wal-Mart
xom	Exxon Mobil Corp

Appendix B

Stata output

Table B.1: Stata output for the final model - estimation and evaluation

```
xtlogit jump r_HF ln_RV ln_Amivest ln_Roll ln_RV_lag1 ln_RV_lag2 ln_RV_lag3
 ln_RV_lag4 ln_RV_lag5 , fe
note: multiple positive outcomes within groups encountered.

Iteration 0:   log likelihood = -29008.949
Iteration 1:   log likelihood =  -28938.22
Iteration 2:   log likelihood = -28938.208
Iteration 3:   log likelihood = -28938.208

Conditional fixed-effects logistic regression   Number of obs     =     49,392
Group variable: share                           Number of groups  =         98

                                                Obs per group:
                                                              min =        504
                                                              avg =      504.0
                                                              max =        504

                                                LR chi2(9)        =    1112.51
Log likelihood  = -28938.208                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        jump |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        r_HF |   .0930559   .0185453     5.02   0.000     .0567078     .129404
       ln_RV |   .4471848   .0144668    30.91   0.000     .4188304    .4755392
  ln_Amivest |   .1668194   .0142939    11.67   0.000     .1388039    .1948348
     ln_Roll |  -.0332221   .0066759    -4.98   0.000    -.0463066   -.0201376
  ln_RV_lag1 |  -.1776523   .0152749   -11.63   0.000    -.2075905   -.1477141
  ln_RV_lag2 |  -.0311889   .0151722    -2.06   0.040    -.0609259   -.0014519
  ln_RV_lag3 |  -.0344457   .0151857    -2.27   0.023    -.0642092   -.0046822
  ln_RV_lag4 |  -.0459234   .0151665    -3.03   0.002    -.0756493   -.0161975
  ln_RV_lag5 |  -.0401386   .0144244    -2.78   0.005      -.06841   -.0118673
------------------------------------------------------------------------------

. estat ic

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
           . |     49,392 -29494.46  -28938.21       9    57894.42   57973.68
-----------------------------------------------------------------------------
```

Table B.2: Stata output for the final model - odds-ratios

```
. xtlogit, or

Conditional fixed-effects logistic regression   Number of obs     =     49,392
Group variable: share                           Number of groups  =         98

                                                Obs per group:
                                                             min =        504
                                                             avg =      504.0
                                                             max =        504

                                                LR chi2(9)        =    1112.51
Log likelihood  = -28938.208                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        jump |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        r_HF |   1.097523   .0203539     5.02   0.000     1.058347     1.13815
       ln_RV |   1.563903   .0226247    30.91   0.000     1.520182    1.608882
  ln_Amivest |   1.181541   .0168888    11.67   0.000     1.148899     1.21511
     ln_Roll |   .9673237   .0064577    -4.98   0.000     .9547492    .9800638
   ln_RV_lag1 |   .8372335   .0127886   -11.63   0.000     .8125397    .8626777
   ln_RV_lag2 |   .9692925   .0147063    -2.06   0.040      .940893    .9985492
   ln_RV_lag3 |   .9661408   .0146716    -2.27   0.023     .9378088    .9953288
   ln_RV_lag4 |   .9551151   .0144858    -3.03   0.002     .9271413    .9839329
   ln_RV_lag5 |   .9606563   .0138569    -2.78   0.005     .9338775    .9882029
------------------------------------------------------------------------------
```

Appendix C

Content of Enclosed Flash Drive

There is a flash drive enclosed to this thesis which contains original and computational data and Gretl and Stata source codes.

- Folder 1: Original data provided by QuantQuote.com
- Folder 2: Computational data generated by VBA and Microsoft Excel
- Folder 3: Stata source codes.