Marius Hallin Ekman
Oskar Eikeseth

Bachelor's thesis

Shorting in the technology sector of the Stock Market

May 2021

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Economics

NTNU
Norwegian University of Science and Technology

Bachelor’s thesis

2021

NTNU

Norwegian University of
Science and Technology

Marius Hallin Ekman
Oskar Eikeseth

Shorting in the technology sector of the Stock Market

Bachelor's thesis
May 2021

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Economics

NTNU
Norwegian University of
Science and Technology

Sammendrag

Ved bruk av Yahoo! Finance har vi hentet paneldata fra 371 teknologifirma oppført ved NASDAQ Composite Index, der vi studerer effekten som fire uavhengige variabler har på andelen aksjer som blir shortet. De fire uavhengige variablene vi valgte å studere er volum, markedsverdi, andelen aksjer holdt av eiere, og andelen aksjer holdt av institusjoner, og vi valgte å bruke short-raten som variabelen for å representere mengden aksjer shortet. En høy short-rate kan gjøre firmaer sårbare for short-klemmer, slik som hendelsen med GameStop i begynnelsen av 2021. Fra analysedelen finner vi en signifikant effekt på short-raten fra andelen aksjer holdt av eiere og andelen aksjer holdt av institusjoner. Vi finner også at volum og markedsverdien har en mer usikker effekt. Disse resultatene bruker vi til å drøfte GameStop-hendelsen, og for å se om det finnes en måte å forutsi, eller til og med unngå, short-klemmer på.

Abstract

By using Yahoo! Finance panel data of 371 technology firms listed at NASDAQ Composite Index, we are researching the effect of four independent variables on the shares. The four independent variables we chose are volume, market cap, shares held by insiders and shares held by institutions, and we chose to use short ratio as the variable for the share of shorted stocks. A high short ratio can make firms vulnerable for Short Squeeze incidents, like the incident with the GameStop stock in the beginning of 2021. From our analysis, we find a significant effect of the percentage of shares held by insiders and the percentage of shares held by institutions on short ratio. Volume and market cap have a more questionable effect. We use these results to discuss the GameStop incident and to see if there is a way to predict, or even avoid short squeezes to happen.

Preface

This assignment is written at the Department of Economics by NTNU, Trondheim.

First off, we want to give a special thanks to our supervisor, Haakon Trønnes, who has been a great help throughout the whole semester. He distributed the dataset for us by coding a file that collected the data from Yahoo! Finance, and gave us important advice and tips for our analytical part. His knowledge has helped us shape this assignment in the way we had visualized beforehand, and for this he really deserves the credit. We also want to thank him for taking the time to read through a draft of the entire assignment and give us feedback on the complete impression of the assignment.

We want to thank Costanza Biavaschi for the lectures in Econometrics, Robin Valenta for the tutorials in Econometrics, and Eli Skaug Rønning for the writing course on BlackBoard. The Econometrics course has been an important part in our analysis, and the writing course has helped us shape our assignment in the way we wanted.

We also want to thank PhD-candidate Musab Khalid Annageeb for taking the time to proofread our assignment before we submitted it. Also, thanks to Oskar`s mentor in econometrics, Christina Rømo, for helpful discussion in issues for our analysis.

And finally, we want to thank the university for good facilitation for Oskar. As a disabled person, taking higher education can present certain challenges. The university and the Economics department have been cooperative in finding solutions on challenges for examinations and other university activities during our thesis writing, as well as the rest of the bachelor program. This is a major reason for our successful completion and submission of the thesis.

Table of contents

1 Introduction

The stock-market is like most other markets, determined by supply and demand. It is a market where one can buy shares in a company they believe in and the better the company does, the higher the demand for their shares will be, leading to an increase in the share price. Therefore, the stock market is also a market with a certain risk. If the company suddenly does worse than expected, the demand for their stocks will decrease and so will their stock-price. The stock market is in many ways an example of a free market, a market with perfect competition with a multiple amount of sellers and multiple amount of buyers.

The stock market is affected by different events, and huge recessions in the market have occurred. Examples are the financial crisis in 2008, stock market crash in 1929, and in newer times, the Covid-19 pandemic closing down a bigger part of the global economy from February/March 2020.

From 2nd of February to 11th of March 2020, the NASDAQ Composite-index decreased with 23,9% (Statista, 2020). Almost ¼ of the total index value was gone in a bit over a month. In situations like this, the yield from shorting could be huge.

On January 28th, 2021, Tesla and SpaceX founder Elon Musk tweeted: “u can’t sell houses u don’t own u can’t sell cars u don’t own but u *can* sell stock u don’t own!? this is bs – shorting is a scam legal only for vestigial reasons” (Musk, 2021). This tweet was published in association with the GameStop incident that happened in the beginning of 2021. For this specific assignment, we were inspired by this tweet and the entire scenario around GameStop.

In January 2021, the GameStop stock price experienced an overwhelming and shocking increase. Mostly driven up by gaming enthusiasts at social media, especially Reddit, coming together to buy stock shares in GameStop. At that time, GameStop was a highly shorted company. Many investors did not have any faith in GameStop's future plans. In the period after this occurred, the GameStop-stock had been violated.

This story has a huge number of different aspects, economically as well as for political sciences. We are issuing the financial part by taking the issue to a higher degree, watching shorting in the technology market and formulated a research question around it.

1.1 Research question

For this thesis, we have chosen to study the following research question:

> "*How does volume, market cap, shares held by insiders and shares held by institutions affect the short ratio in the Stock Market?*"

We originally wanted to look at the amount of shorted stocks and use this as our dependent variable. But, we quickly realized that we needed to use a variable that could represent the shares shorted compared to the company's size, to exclude the effect of bigger companies generally having more activity than smaller companies. We then chose to use the short ratio as our dependent variable, which tells us how many shares that are shorted compared to the company's average daily trading volume.

With this assignment, we hope to find out if the share of shorted stocks is random, or if variables within the Stock Market have an influence on it. To delimit our assignment, we have chosen to study the effect of only four variables, so that the assignment is manageable to finish within our timeframe. To study these variables impact on shorting, we are going to use the OLS regression method.

Further on in this assignment, we are also going to look closer at the GameStop incident and want to discuss what happened there. We want to address the incident, and hopefully find some guidelines on how we can prevent or predict such an incident in the future.

1.2 Background

The GameStop incident was a recent event that happened only weeks before the deadline for submitting our research question was set. This was an incident that garnered a lot of attention in the public, and exactly what happened is something we are going to discuss later in this assignment. This incident was the main factor that made us want to choose this certain research question, and the tweet by Elon Musk which we introduced earlier made it even more interesting for us. It made us curious as to how the stock market can be “manipulated” in certain ways, while it also made some people earn a lot of money while others lost a lot. We found this interesting, and it helped us make our minds up when formulating our research question.

When we read about the GameStop incident, we felt like it had similarities with the commonly used term David vs Goliath. This term is used for explaining when a smaller or less resourced part (David) faces a more resourced part (Goliath). Goliath is expected to win, or at least get better off, while David is expected to lose. Yet, this is not the outcome, and it turns out that by using the limited resources in a smart manner, David manages to beat Goliath. In our case, Goliath represents the big investor companies while David represents everyday-people investing a bit of their savings into stocks without any power in the market. In this case with GameStop, the small-investors managed to push the stock-price up and some of them ended up with huge gains. This also resulted in enormous losses for big investors who were shorting the stock.

Exactly how this incident happened and how the small investors were able to push the price up, is something we will discuss closer in chapter 2 of this assignment.

1.3 What is shorting?

The way most of us are familiar with the Stock Market, is how one can buy and sell different companies’ shares and hopefully make some money in the process. The sum of shares is referred to as a stock. The share price is often volatile, which means that the share price will experience oscillations. After a person purchases a share, they hope that the share price rises so that they can sell the share for a higher price than they bought it for and make a profit that way. If they are unlucky with the share purchase, the share price may fall, and they could be forced to sell it with a loss. The difference in share price is what a person yields when purchasing a share. If the share price is 10 when a person purchases the share, it really depends on what the share price is when the person wants to sell. If the price is 8, the person experiences a loss of 2 per share. If the price is 12, the person experiences a gain of 2 per share.

Some people that are highly involved with the Stock Market, can learn ways to predict if certain shares will rise or fall in price. An example of a man that has had a lot of success in the Stock Market, is Warren Buffett (1930-). It is obviously never completely safe to trust their predictions, but they can get it right in most cases like Buffett has done. They will then choose to buy the shares that they predict will rise in price. Yet, there is also a way to make money on the shares they predict will experience a fall in the price. This is where the term shorting comes into play.

To illustrate with an example, let's say you borrow a friend's phone that you do not own. You know that a new model will be released in just a few weeks, so your friend’s phone will decrease in value. You therefore choose to sell your friend’s phone now, so that when the new model is released, you can buy an identical phone back for a lower price. You will then give your friend back the phone that you bought, and you keep the difference in money that you gained when flipping the phones. This is basically the same as what you are doing when shorting a share. The only assumption we need to make for this example, is that the phone you borrow and sell, and the phone you later buy back, are completely identical. It will then be indifferent for your friend which phone you return.

In more theoretical terms, there are a number of ways to achieve a short position, where the most fundamental method is the "physical" short selling. This involves borrowing assets such as shares or bonds and selling them. The investor, or borrower, will later purchase the same number of the same type of share or bond, in order to return them to the lender. If the price has fallen in the meantime, the investor will have made a profit equal to the difference in price. Conversely, if the price has risen then the investor will bear a loss (*Short (finance)*, 2021). This is a way of trading stocks that is extremely risky, and we will dig closer in on why later in this assignment.

1.4 Hypothesis

From part 1.1 where we introduce our research question, it can be seen that we want to study four different independent variables' effect, on the company's relative amount of shorted stocks compared to its size. We want to use the company's volume to measure their size. Therefore, we chose to use the short ratio as our dependent variable, or y-variable in a statistical manner, since it includes volume in its formula. The four independent variables for the model are volume, market cap, shares held by insiders, and shares held by institutions.

For every one of the independent variables, we have formulated a hypothesis. By each hypothesis, we also have a null hypothesis. The hypothesis indicates a negative effect of x, the independent variable, on y, the dependent variable. The null hypothesis indicates that this is not the case. Our main hypothesis states the following:

H_1: The volume of stocks has a negative effect on short ratio.

H_2: The marked cap has a negative effect on short ratio.

H_3: Shares held by insiders has a negative effect on short ratio.

H_4: Shares held by institutions has a negative effect on short ratio.

Our null hypothesis will then be:

H_0: not H_i, where i range from 1 to 4.

Instead of writing a specific H_0 for each of the hypotheses, we choose to just save us some trouble by using i to indicate that the null hypothesis is the same for every hypothesis.

These hypotheses will be useful when we look at the results of the models and when we conduct our analysis and conclusion. Intuitively, we have reasons to believe the variables will have a negative effect. We believe a high volume will have a negative effect, due to the short ratio mathematically being calculated with volume as a parameter in the denominator of the fraction. Therefore, it makes sense that when volume increases, the short ratio should increase (given that the numerator is held unchanged). For market cap, we believe that when the market cap is high, people are less willing to short the company's stock due to the expectation of bigger companies experiencing a positive evolvement. For shares held by insiders, we expect the short ratio to decrease when there is a higher share of stocks held by insiders. This is due to the fact that when insiders have a stake in the company themselves, they will have an incentive to work harder to increase its value so that they gain money. As an investor, you therefore have reasons to believe that the company's stock price will increase due to increased profits, and you will not short it. For shares held by institutions, we look at it as an indicator of professional investors expecting the company's stock price to increase. If the share held by institutions is high, a lot of professional investors have chosen to invest in it on behalf of institutions. As an investor you therefore have reasons to believe that the company's stock price will increase, and you will not short it.

When we have ran our MLR model, we will run a test of the hypothesis to see which hypothesis best fits the sample. We can then, at a certain significance level, conclude whether we can reject the null hypothesis or not. More in depth analysis of how we interpret this, is something we will get back to in chapter 5.

2 Theory

To understand some of the basics of the stock market, shorting and the GameStop incident, we want to use this part of the assignment to address some fundamental knowledge about these subjects. This will help to further understand why we chose to study shorting for this thesis, and also help in explaining some of the terms that later will be used in the analytical part of the assignment.

2.1 Summary of published research

We are not the first ones to write about shorting of stocks. Several papers have been published, and we now want to mention some of the ones we found most interesting considering our thesis. We have used these articles actively in our thesis, to try and get a wider perspective on the subject.

In Laurent Cohen, Karl B. Diether and Christopher J. Malloy's paper, they found the following result: "average abnormal returns for stocks experiencing an outward shift in shorting demand are −2.98% in the following month". They also comment on their results on other variables that is interesting for this assignment, like for instance institutional ownership; "both institutional ownership, which proxies for ease of shorting a stock, and volume, which could proxy for a number of effects including recall risk, disagreement, and liquidity, do not significantly affect abnormal returns after controlling for the shifts." (Cohen, Diether & Malloy, 2007). Cohen, Diether and Malloy try to determine both supply and demand. They concluded that there is a huge effect of shorting demand on future returns, but shorting demand is not related to private information.

Itamar Drechsler and Qingyi Freda Drechsler go deeper inside premium shorting in a paper from 2014. They used the interesting variable SIRIO, denoting Short Interest Ratio relative to Institutional ownership as a proxy for shorting demand. SIRIO will later be the same as our dependent variable Short Ratio. The paper states that "l. Sorting stocks into deciles using SIRIO, we find a large and statistically significant average return of 1.48% per month on the corresponding CME portfolio, with a FF4 alpha of 1.54% per month." (Drechsler & Drechsler, 2014)

Even though the Gamestop short squeeze happened recently, Dr. Usman Cohen has published a paper about this incident in January 2021. This paper is commented on more widely in part 2.4 about the GameStop Incident. It is difficult to predict the long-term consequences of this incident, but we got a good picture of the incident, what was happening and why it happened. Chohan examines some counter-hegemonic elements underlying for the incident (Chohan, 2021).

2.2 Opportunity cost of investments

Today, many will consider the stock market to be a great investment for savings that do not need to be liquidated in the nearest future. This is due to the possibility of getting a high return and increasing the savings. Yet, there are also a number of different ways to invest the savings to achieve a return, and we want to address the opportunity cost of investing in the stock market. To do this, we are going to break it down into five different ways of investing savings/the leftover money one has after paying all necessary bills. At the end of the day, where a person invests their money, often depends on the person's attitude towards risk. Are they risk averse, risk-seeking or risk neutral?

If the person is risk averse, they prefer lower risks and accept a lower, but safer, return. A risk averse person will often choose to keep their savings in the bank, where the return is extremely low, but the money is completely safe and liquid. They may also choose to invest in funds which have a bit higher return, and still are safe, but where their money is not liquidated if they suddenly need it. These are two of the ways to invest the savings.

If the person is risk-seeking, they prefer a higher return and do not mind the risk that comes along with it. These people may choose to invest in the stock market, where the share price is often very volatile. This is the third way to invest. By investing in the stock market, it is difficult to conclude to which extent a person is risk-seeking simply by this choice. It really depends on which companies they invest in. Some companies have a slow and steady growth over time, which could make them a quite safe investment. Others experience lots of oscillations daily and are harder to predict. Day-traders are often risk-seeking and take advantage of shares with lots of oscillations to try and make a quick profit. Shorting will also be placed in this category, which is for people who are risk-seeking. We can look at shorting as the inverse alternative to investing in stocks. This means we can assume that investors will purchase if they believe the stock price will rise, and short sell if they believe the stock is overvalued and will decrease.

For the two final points on how the investor can use their money, it does not depend on their attitude towards risk, but rather their preferences. The fourth way to invest, is in other value papers such as bonds for instance, or to invest in property. This point is quite broad and we will not look too close at it. You can be either risk-seeking, risk averse or risk neutral, and still invest in such, because the extend of risk will vary in each individual case. Hence, it does not really depend on the investors attitude towards risk, but rather their preferences and expectations. The fifth and final way is to consume the money on goods or services by purchasing something. This can be items such as a TV, a car, or a vacation. In this case, the investors attitude towards risk is also not the determining factor, but rather their preferences on how they want to live their lives.

These options are opportunities for using money. For each dollar (or any other currency) you have, you are theoretically standing in front of this investment dilemma. On the other hand, in real life most people will choose a combination of these. The kind of combination people chose depends on their preferences and characteristics such as the amount of money they have, their attitude towards risk, if they are time consistent, their needs of liquidity and more. In fact, for every choice you make, there will be an opportunity cost. The opportunity cost will be the best alternative way to place the money.

2.3 Shorting versus investing

In the introduction part, we briefly discussed what shorting is and one of the ways to achieve a short position. Further on, we want to look at the mathematical aspect of shorting to illustrate the risk versus reward. Also, we want to introduce another way to achieve a short position, which is less risky than short selling.

To mathematically illustrate how investing in stocks and shorting works, we can give the chance for the stock to increase its value, the variable q. This will be a number between 0 and 1, due to its probability of increasing being between 0% and 100%. Hence, by multiplying this number by a hundred, we get the percentage value for the chance of the stock to increase in value. Inverse, this means that the chance for the stock to decrease in value is 1-q, given that the stock is volatile and therefore is expected to either rise or fall, but not stay constant. Mathematically, this can then be written as:

Probability of increase or decrease $= q + (1-q) = 1$

We denote positive yield from when the stock increases, as the variable p. We denote the negative yield from when the stock decreases, as n. We define p and n as:

$p>0, \quad 0<n<1$

The reason p and n are defined differently, is because the stock can increase more than a hundred percent of its current value (hence, p can be anything above zero), but can only decrease up to a hundred percent of the stocks value before reaching zero (hence, it is capped at 1, or 100%). We can then define the expected value of a stock as:

$E(\text{stock}) = q(1+p) + (1-q)(1-n)$

(Note that because we want to use the yield as a multiplier, we need to add 1 to the positive yield and subtract negative yield from 1 to make it mathematically correct. This makes it so that if we experience a 10%, or 0,1 increase, we get the multiplier 1,1 for positive yield. If we experience a 0,1 decrease, we get the multiplier 0,9 for negative yield.)

We cannot calculate a general expected value from this equation, due to q, p and n being different at different times, and also different for every stock in the market. These variables are therefore unknown to us. This means that we cannot predict how a stock will develop using only this equation, because there will be a lot of other factors that affect q, p and n, such as for instance the company’s annual results, news in the media, or shocks in the stock market. Regardless, what we can use this equation for, is to illustrate the risk of an investment compared to short selling. For the first part of the equation, we have the expected positive yield of the stock. For the second part, we have the expected negative yield. If we add the investment to the equation, as the variable I, we can illustrate the expected return of the investment as:

$$E(return) = q(1+p)I + (1-q)(1-n)I$$

Note that we still cannot calculate a result from this equation, due to the variables q, p and n being unknown. But, since we have defined p as being anything above 0, while n is limited within the interval 0 to 1, with everything else being equal, we can expect the absolute value for the two parts of the equation to be:

$$|q(1+p)| > |(1-q)(1-n)|$$

This means that the expected absolute value of positive yield is greater than the expected absolute value of negative yield. Intuitively, this makes sense due to how p and n are defined. By investing in the stock, the positive yield can be up to unlimited due to the stock price being able to increase without restrictions, while the negative yield can only be as much as the original investment. Because short selling can be seen as the inverse alternative to purchasing, this means that the roles switch. We will then gain from negative yield and lose from positive yield. Hence, as seen from the equation, we can only gain up to 100% of the stock price per share, while the possible loss can be as much as up to unlimited. The loss really just depends on how much the stock increases in value. We can therefore see that short selling contains a much greater risk than investing traditionally, because there is a chance for such great losses.

Another alternative way to achieve a short position rather than short selling, is through investing in an inverse exchange traded product, also known as an ETP (Nordnet, 2020). It provides a positive return when the market that is being tracked, falls. For example, if the index or asset that is being tracked falls by 1%, a 3x Short Daily ETP would rise by 3% on that day, and the investor would experience a 3% gain. If the index or asset being tracked rises by 1%, a 3x Short Daily ETP would fall by 3% on that day, and the investor would experience a 3% loss. The reason shorting through ETPs is a lot safer than short selling, is because losses are capped at the original investment. This means that shorting through ETPs contain the same risk as purchasing stocks in a traditional manner and does not contain the same great risk as we illustrated above (*Why use short ETPs to achieve a short position?*, 2017).

By studying the theory around risk preferences, we can create a picture of how investors, generally speaking will invest in the finance market. It is reasonable to assume that risk-seeking investors will look at shorting as more attractive than risk averse investors do. It is also important to remember that you cannot say in an absolute term whether a person or investor is risk-seeking or risk averse. There is a fine line between being risk averse, risk neutral or risk-seeking, and a person can act differently in different situations. Another important aspect to also remember, is that risk theory is not the only factor that affects an investor's decision on whether to short or not. There are several other factors that have a say in that decision, and this is the reason we chose our research question. We want to take a closer look at some of the other factors that are harder to evaluate and see whether they actually have an effect on this decision or not. We therefore chose volume, market cap, shares held by insiders and shares held by institutions as our four main factors to study.

2.4 The GameStop Incident

To begin this part where we explain the GameStop incident, we can start off with introducing what GameStop actually is. GameStop is an American technology company, more correctly a high street shop, which is known worldwide and sells games, consoles, and other electronic devices (Grant, 2021). Their stocks, GME, can be traded at the New York Stock Exchange (NYSE), as well as several other stock exchanges around the world.

The GameStop stock was a highly shorted stock, where the hedge fund group Melvin Capital was one of the major backers of the shorting. When this group started shorting GME, the stock price was around 15 USD, and in the faith of further decreasing in price. The reason was, amongst other things, problems with adaptation to digitalization and lockdown restrictions that made it difficult and less attractive to seek out to physical stores (Chohan, 2021). We can also assume that the risk of being infected by the COVID-19 virus from going out in public, made more people prefer online shopping and give a kind of correlative effect.

As mentioned earlier, the social media platform Reddit was highly involved in raising the price, by using the group r/wallstreetsbets, from now on referred to as WSB. This group has had an increase in members since the COVID-19 pandemic started (Kolhatkar, 2021). Also, the well-known Tesla-founder Elon Musk commented on the social media platform Twitter, trying to “fire up” the members in WSB and others interested.

GME increased and reached their highest stock price at 347 USD. That resulted in a loss for Melvin Capital which needed to borrow around 3 billion USD bailout from the government to cover their losses (Kolhatkar, 2021).

By January 22nd. 2021, the GME were shorted by over 140%. That was an opportunity for the small investors from WSB and others to do something. By January 27th, the stock price exploded by more than a 700% increase. When the WSB was “done” ruining for short sellers of GME, they ran away to other stocks. Such as Nokia, Blackberry, Koss Corporation and Eastman Kodak (Chohan, 2021). It was also speculated that something similar could happen with another physical business called AMC. Which has also suffered during COVID-19 and been shorted by Wall Street traders (Kolhatkar, 2021).

These types of incidents, such as what happened to GameStop, is called a short squeeze. A short squeeze is characterized by when a stock, or another asset, jumps sharply high in price. That again forces the short sellers, who is betting the price will fall, to buy shares in order to forestall an even greater loss (Mitchell, 2021). As we can see, this is a self-reinforcing effect on the price that makes it rise even more, and makes the losses for short-sellers even greater.

3 Methodology

Let us now introduce the methodology we are going to use, to try and explain whether or not volume, market cap, shares held by insiders and shares held by institutions actually have a negative effect on short ratio or not. If they do have an effect, we also want to try to explain how big this effect is. To analyse this effect, we are going to be using a model named ordinary least squares linear regression model and we will now introduce this model and how it works. We will also introduce the dataset we are using for our analysis and how our variables are calculated.

3.1 Ordinary Least Squares

To study our research question, we find it suitable to use a linear regression model to explain the relationship between our four explanatory variables volume, market cap, shares held by insiders and shares held by institutions, and our explained variable short ratio. A linear regression model can be categorized into being either a simple linear regression model (SLR) with only one explanatory variable, or a multiple linear regression model (MLR) with more than one explanatory variable. Since we have more than one explanatory variable, we want to study the effect of, we will mainly focus on using the MLR in this assignment. We will use Stata to run our regression analysis and interpret the results found in Stata to try to explain this relationship.

In a linear regression, the relationships between the explanatory variables and the explained variable are modelled using linear predictor function, whose unknown model parameters are estimated from the data. The linear regression model we are going to be using in this assignment chooses the parameters of a linear function, of a set of explanatory variables by the principle of least squares (*Linear regression*, 2021). This model is called the Ordinary Least Squares (OLS) linear regression model. This means that it minimizes the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset, and those predicted by the linear function of the independent variable (*Ordinary least squares*, 2021). The predictor function for the OLS when running a simple linear regression is called the population regression function, and can be expressed as:

$y_i = \beta_0 + \beta_1 x_i + u_i$

and while running a multiple linear regression, it can be expressed as:

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$

where i indicates a particular observation and we have k number of explanatory variables. The y is data on the explained variable, while x_1 is data on the explanatory variable 1, x_2 is data on the explanatory variable 2, and so on for all k variables. The u_i is the error term in the model, and captures everything that affects the explained variable, which the explanatory variables do not capture. The betas, β, will tell us how much a change in the explanatory variable x, will change the explained variable, y. The β_0 is a constant, which tells us what the model estimates the explained variable to be when every explanatory variable is equal to zero.

In the model, we assume that such a relationship as the linear population regression function is true in the population. When we run the regression, we use the same function but replaces y with ŷ, β with β^, and u with û. This function is called the sample regression function. When we use the variables without the hat, it means that we have the true parameters. Since we want to predict the variables and they therefore are unknown to us, we use the hat to indicate that these are estimates and not the true parameters.

When we run the regression analysis, we will use a dataset and determine which variable in the dataset we want to explain, and which variables we want to use as explanatory variables. When we run the regression in Stata, we will get an output with the results which we can interpret. It will tell us the constant, the betas for the different explanatory variables, how much these explanatory variables actually explain the explained variable, and more. Exactly how we interpret these results we get back to later in the analytical part of the assignment when we run our own regression. By using Stata to get these parameters, we save ourselves from a lot of work. But note that there are formulas which makes us able to calculate these parameters by hand as well.

For the model to be unbiased, we have a set of assumptions which needs to hold. If any of these assumptions break, it means that the results we get from the model are biased. For the simple linear regression model, we have five assumptions, SLR.1 to SLR.5. They sound the following:

> SLR. 1: Linearity in the model.
>
> SLR. 2: We have a random sampling.
>
> SLR. 3: Enough variation in x, the explanatory variable, and no perfect collinearity.
>
> SLR. 4: Zero conditional mean, which implies that the regressor x should not be correlated with the error term. Can also be written mathematically as: $E(u_i|x_i) = 0 \rightarrow Cov(x_i, u_i) = 0$
>
> SLR. 5: Homoscedastic. The variance in the error term is homoscedastic, which means that it stays the same. This means that as x increases, the error term won't change. This can mathematically be written as $V(u_i|x_i) = \sigma^2$, which is constant.

For the multiple linear regression model, we have six assumptions and not just five. These are MLR. 1 to MLR. 6, where MLR. 1 to MLR. 5 are the same as SLR. 1 to SLR. 5 and are usually called the Gauss Markov Assumptions. MLR. 6 sound the following:

> MLR. 6: Normality. The population error is independent of the explanatory variables and is Normally distributed with zero mean and variance of σ^2:
>
> $u \sim \text{Normal}(0, \sigma^2)$

MLR. 6 is the strictest and strongest MLR assumption. MLR. 6 effectively encompasses the two other MLR assumptions pertaining to the error term. That is, MLR. 4 and MLR. 5, ”Zero Conditional Mean” and ”Homoskedasticity” respectively. MLR.6 encompasses these because it would not be possible to have a normally distributed error term, if the error term were correlated with any X's, either in the error terms mean value or in the error terms variance. In other words, if MLR. 4 or MLR. 5 does not hold, neither will MLR. 6. MLR.1 through MLR.6 are called the ”Classic Linear Model” (CLM) assumptions.

When we have ran the regression model and the assumptions hold, we can start interpreting the results. While doing this, we can run t-tests, f-tests and control for p-values. A t-test is also often called Students t-test and is a hypothesis test based on the Students t-distribution. It is used to test if the mean value in a normally distributed dataset is significantly different from a null hypothesis, if there are significantly differences between the mean value in two datasets, or if the slope of a regression line is significantly different from zero (*T-test*, 2018). The f-test on the other hand, is a test where the test statistic has an F-distribution under the null hypothesis. It is used when we want to compare different regression models to identify which one fits the population from which the data were sampled (*F-test*, 2021). And finally, the p-value is the probability of obtaining test results as extreme as the results actually observed, under the assumption that the null hypothesis is correct. A small p-value therefore indicates that such an extreme observed outcome would be very unlikely under the null hypothesis, and we can therefore reject the null hypothesis. If the p-value is large, this means that there is a great chance that we observed such an extreme observation under the assumption of the null hypothesis, and we therefore fail to reject it (*P-value*, 2021). We have given critical values we can use when testing for hypotheses, which we can use to check if we can reject the null hypothesis under a certain significance level. There is a lot that can be said around this, but we choose to just briefly introduce the terms in this part of the assignment, and rather look closer at them under the analysis part.

3.2 Variables and dataset

The data used for this assignment are collected from Yahoo Finance. It is in total 371 observations, in other words companies. All the companies are technology companies listed at the NYSE and NASDAQ Composite within the technology sector. The data on Yahoo! Finance originally contained 443 companies, but some of them lacked data on certain variables. Because of this, we chose to use a clean shaved version, and we still reach a true picture of the market without any sample size issues. The full dataset that we have used is available in the Appendix.

Yahoo! Finance is a website publishing financial news, data and commentary. Yahoo! Finance is one of the most credible pages for financial data.

We have chosen to use the Short Ratio, which some choose to call the Short Interest Ratio, as our dependent which we want to explain. For traders, the short interest can be helpful in describing reactions in the stock for bearish or bullish sentiment. If the short interest is high, the stock could be a candidate for a short squeeze (Duggan, 2019). Something that brings us back to the GME-incident which we explained earlier. The formula can be written as $\frac{SI}{ADTV}$, where SI denotes Short Interest and ADTV denotes Average Daily Trading Volume (Chen, 2021). From now on, Short Interest will be referred to as “Shares Shorted” and Average Daily Trading Volume will be referred to as “Volume”. The short ratio is directly a consequence of these two variables. If Shares Shorted increases, all else equal, this ratio will increase. While if Volume increases and Shares Shorted remains the same, the ratio will decrease.

We are using 4 different variables as independent, explanatory variables. We now want to present these variables, along with their definitions:

Volume shows the actual number of shares that are traded in a certain period of time, often referred to as “daily volume”, where the time period equals one day of trading. The volume gives an indicator used to measure the relative significance of a market move (Hayes & Anderson, 2021). For investors, volume is an important parameter because it tells something about the market activity and the liquidity of the stock. A higher volume will obviously make the stock more liquid since it will be easier to sell the stocks for cash.

Market Cap stands for market capitalization and gives us the total market value of a company's outstanding shares of stock (Fernando & Boyle, 2021). The formula for the market cap can be written as: share price * shares outstanding. Intuitively enough this needs to be the total value, since share outstanding is the total number of shares and every share has a certain value, the share price. So the total value will be the value for all the shares summarized. So, if you are interested in buying a whole company listed at a stock exchange, the market cap will be the company’s worth and most likely the price you must pay.

Shares held by insiders refer to the percentage of shares that are held by directors, executives, and senior officers within the company. This variable is interesting because it will give us a sign how a big part of the stocks is held by insiders, that might have biases in buying their own stocks, or are obligated by their position to hold a certain share of stocks. The formula to calculate the share held by insiders can be written as: $\frac{Shares\ held\ by\ insiders}{Shares\ outstanding}$ (*Shares Owned By Insiders*, 2021).

Shares held by institutions refers to the percentage of shares that are held by investment banks, pension funds, insurance companies, mutual funds, hedge funds, etc. This variable gives us interesting information about the market, because this is institutions knowing what they are doing and they are the largest holder of stocks, as more or less professional investors (Jagerson, 2017). The formula is defined as $\frac{\text{Shares owned by institutional investors}}{\text{Shares Outstanding}}$ (*Shares Owned By Insiders*, 2021). Institutions normally hold their positions for a longer period of time, due to their investment decisions being thoroughly analysed in analytical research (Kenton, 2020).

We have also run an alternative model where we change the volume and market cap variables to their logarithmic form. This will make the variables impact measure in percentage instead of their respective units. The variables are in every other way equal to the earlier variables. The results in this model are presented in chapter 4.4.

4 Results

In this part of the assignment, we are going to use the dataset in Stata and present the results we obtained. We are going to get some descriptive statistics from using the summarize command on our explanatory variables and our explained variable, which is going to be the number of observations, the mean value, the standard deviation, the minimum value observed in the dataset, and the maximum value observed in the dataset, for each of the variables. We are also going to be running a single linear regression model for each of the explanatory variables on the explained variable, ShortRatio, and look at what results we get. Then, we will finally run a multiple linear regression model with all the explanatory variables and see whether the results will change compared to the single linear regression models. We are only going to be presenting the data found in this part, while we are going to use chapter 5 to analyse and interpret the results.

4.1 Descriptive statistics

As mentioned in the introduction for this chapter, the variables we are going to be looking at as the descriptive statistic is the number of observations, the mean value, the standard deviation, the minimum value observed, and the maximum value observed, for each of the variables. All the data were collected on February 11th, 2021, and the variables therefore represents what the value were at that point in time. We have made a table with the results from Stata, which gives a clear overview of the results. The table is presented below:

Variable	Observations	Mean	Std. Dev.	Min value	Max value
MarketCap	371	2,84e+10	1,46e+11	6,32e+07	2,10e+12
Volume	371	2 222 400	7 197 681	9	9,74e+07
HeldByInsiders	371	0,0790022	0,1324656	0	0,8965
HeldByInstitutions	371	0,8334825	0,2011884	0,008	1,1672
ShortRatio	371	3,868086	2,719206	0,11	16,03

As we can see, we have 371 observations for each variable. The reason for this is that we rinsed the dataset from observations that lacked data for the variables we chose to analyse. The dataset from Yahoo Finance originally contained 443 technology firms, which means we had to remove 72 incomplete observations. Therefore, the 371 observations that we kept in the dataset, all contains the information and data needed for our analysis.

For the MarketCap variable, which represents the firms market cap, it has a mean value of 28 400 000 000 US dollars (28,4 billion), while the minimum market cap observed is 63 200 000 US dollars (63,2 million) and the maximum market cap observed is 2 100 000 000 000 US dollars (2,1 trillion). As we can see from this, there is quite a gap from the technology firms with the lowest market cap and the firms with the highest. The standard deviation for this variable is calculated to be 7 197 681 US dollars.

For the Volume variable, which represents the firms average daily volume at the time we collected the dataset, the mean value for daily volume were 2 222 400 number of shares. The minimum average volume observed is 9 shares, while the maximum average volume observed is 97 400 000 shares. The standard deviation for this variable is calculated to be 7 197 681 shares. As seen, the standard deviation for this variable is actually higher than the variables mean value. This can seem strange but can be explained due to a few companies having an extremely high volume, compared to the majority of the companies. This can be seen in the scatterplot for volume and short ratio, which is found in table 5.1 in the Appendix. When most of the companies have a low volume, the mean value will be low. When a few companies then have such a high volume, the big differences in volume will increase the standard deviation to such an extent that its actually bigger than the mean.

For the HeldByInsiders variable, which represents the percentage of the firms outstanding shares which is held by insiders within the firm, the mean value is 0,0790022. This can be converted into a percentage, by multiplying with a hundred. We will then get the mean value, which is approximately 7,9%. The minimum value observed in the dataset is zero, while the maximum value observed is 0,8965, which represents 89,65% of the outstanding shares. The standard deviation for this variable is calculated to be 0,1324656, or 13,24656%.

For the HeldByInstitutions variable, which represents the percentage of outstanding shares that is held by institutions, the mean share held by insiders is estimated to be 0,8334825, or 83,34825%. The minimum value observed is 0,008, or 0,8%, and the maximum value observed is 1,1672, or 116,72%. Intuitively, it does not make sense that anyone can hold a larger number of shares than the firm have outstanding shares. Yet, the reason we observe this can be explained as due to errors in the data that is a result of slow updating. There can be delays in updating publicly available data, so the figures released in an institution's report correspond to an institutional holding's date. These dates generally differ somewhat among all of the institutions that hold a company's stock, resulting in differences that could impact the reported percentage for total institutional holdings being displayed (Loth, 2019). The standard deviation for this variable is estimated to be 0,2011884, or 20,11884%.

And finally, for the ShortRatio variable, which is our explained variable in the model and represents the short ratio, has a mean value that is estimated to be 3,868086, with a minimum value observed of 0,11 and a maximum value observed of 16,03. The standard deviation for this variable is estimated to be 2,719206.

4.2 Single Linear Regression model (SLR)

For this part of the assignment, we want to run a single linear regression (SLR) model for all our explanatory variables, on behalf of the explained variable, the short ratio. We will present the results found in these models, so that we can later compare the results of the SLR models with the results from the multiple linear regression (MLR) model. We have converted the results found in Stata into an Excel-table, which we have included for each explanatory variable. We also wanted to include a scatter plot for each of the explanatory variables with short ratio, to illustrate the distribution and the spread of the observations compared with the fitted sample regression function line. These scatterplots can be found in table 5.1 to 5.4 in the Appendix.

For the Volume variable, the SLR model is estimated as the following:

ShortRatio	Coef.	Std. Err	t	P>|t|	[95% Conf. Interval]	
Volume	-5,06e-08	1,94E-08	-2,88	0,004	-9,42e-08	-1,77e-08
Constant	3,992	0,1463	27,28	0,000	3,7047	4,2802
					Obs	371
					R-squared	0,0219

As we can see from the table above, we have a coefficient of -0.0000000506 for volume. This looks like quite a small number at first glance, and how we interpret it is something we will come back to in the analysis part. The constant is the beta zero in the model and is estimated to be 3,992. Hence, the sample regression function for this model can be written as:

$ShortRatio_i = 3{,}992 - 0.0000000506 * volume_i$

As seen from the sample regression function, if the volume is zero, the model estimates the short ratio to be 3,992.

For the volume coefficient, we have the following estimations:

The standard error is estimated to be 0.00000194. The t-value is -2,88 and the p-value is 0,004, and we will get back to what these values tells us in the analysis. The model also estimates that at a 95% confidence interval, the true beta parameter will be within the interval -0.00000942 and -0.00000177.

For the constant, we have the following estimations:

The model estimates the standard error to be 0,1463, the t-value to be 27,28 and the p-value to be 0. At a 95% confidence interval, the true beta parameter is estimated to be within the interval 3,7047 and 4,2802.

And for the last variable we want to look at, we have the R-squared which is estimated to be 0,0219 and tells us that the model estimates volume to explain 2,19% of the variation within the short ratio variable.

For the MarketCap variable, the SLR model is estimated as the following:

ShortRatio	Coef.	Std. Err	t	P>|t|	[95% Conf. Interval]	
MarketCap	-2,28E-12	9,61E-13	-2,37	0,018	-4,17E-12	-3,88E-13
Constant	3,933	0,1429	27,52	0,000	3,6517	4,2138
					Obs	371
					R-squared	0,015

As we can see from the table above, we have a coefficient of -0.00000000000228 for market cap. The constant is the beta zero in the model and is estimated to be 3,933. Hence, the sample regression function for this model can be written as:

$ShortRatio_i = 3,933 - 0.00000000000228 * MarketCap_i$

As seen from the sample regression function, if market cap is zero, the model estimates the short ratio to be 3,933. The R-squared in the model is estimated to be 0,015, which tells us that the model estimates market cap to explain 1,5% of the variation within the short ratio variable.

For the market cap coefficient, we have the following estimations:

The standard error is estimated to be 0.000000000000961, the t-value is -2,37 and the p-value is 0,018. The model also estimates that at a 95% confidence interval, the true beta parameter will be within the interval -0.00000000000417 and -0.000000000000388.

For the constant, we have the following estimations:

The model estimates the standard error to be 0,1429, the t-value to be 27,52 and the p-value to be 0. At a 95% confidence interval, the true beta parameter is estimated to be within the interval 3,6517 and 4,2138.

For the HeldByInsiders variable, the SLR model is estimated as the following:

ShortRatio	Coef.	Std. Err	t	P>|t|	[95% Conf. Interval]	
HeldByInsiders	0,2309	1,0685	0,22	0,829	-1,8703	2,3321
Constant	3,8498	0,1646	23,38	0,000	3,526	4,1636
					Obs	371
					R-squared	0,0001

As we can see from the table above, we have a coefficient of 0,2309 for held by insiders. The constant is the beta zero in the model and is estimated to be 3,8498. Hence, the sample regression function for this model can be written as:

$ShortRatio_i = 3{,}8498 + 0.2309 * HeldByInsiders_i$

As seen from the sample regression function, if shares held by insiders is zero, the model estimates the short ratio to be 3,8498. For the R-squared, we have that the model estimates it to be 0,0001, which tells us that the model estimates the shares held by insiders to explain 0,01% of the variation within the short ratio variable.

For the coefficient of the shares held by insiders, we have the following estimations:

The standard error is estimated to be 1,0685, the t-value is 0,22 and the p-value is 0,829. The model also estimates that at a 95% confidence interval, the true beta parameter will be within the interval -1,8703 and 2,3321.

For the constant, we have the following estimations:

The model estimates the standard error to be 0,1646, the t-value to be 23,38 and the p-value to be 0. At a 95% confidence interval, the true beta parameter is estimated to be within the interval 3,526 and 4,1636.

For the HeldByInstitutions variable, the SLR model is estimated as the following:

ShortRatio	Coef.	Std. Err	t	P>|t|	[95% Conf. Interval]	
HeldByInstitutions	3,9671	0,6726	5,90	0,000	2,6446	5,2898
Constant	0,5615	0,5767	0,97	0,331	-0,5724	1,6955
					Obs	371
					R-squared	0,0862

As we can see from the table above, we have a coefficient of 3,9671 for held by institutions. The constant is the beta zero in the model and is estimated to be 0,5615. Hence, the sample regression function for this model can be written as:

$ShortRatio_i = 0,5615 + 3,9671 * HeldByInstitutions_i$

As seen from the sample regression function, if shares held by institutions is zero, the model estimates the short ratio to be 0,5615. For the R-squared, we have that the model estimates it to be 0,0862, which tells us that the model estimates the shares held by institutions to explain 8,62% of the variation within the short ratio variable.

For the held by institutions coefficient, we have the following estimations:

The standard error is estimated to be 0,6726, the t-value is 5,90 and the p-value is 0. The model also estimates that at a 95% confidence interval, the true beta parameter will be within the interval 2,6446 and 5,2898.

For the constant, we have the following estimations:

The model estimates the standard error to be 0,5767, the t-value to be 0,97 and the p-value to be 0,331. At a 95% confidence interval, the true beta parameter is estimated to be within the interval -0,5724 and 1,6955.

4.3 Multiple Linear Regression model (MLR)

Now we want to run a multiple linear regression (MLR) model, where we include all our explanatory variables in the same model. We find it interesting to do it this way, to show how the coefficients for the variables may vary once we add more variables to the model. For this part, we have done the same as for the SLR models. We have run the model in Stata, and made an Excel-table out of the results found there. The table is presented below:

ShortRatio	Coef.	Std. Err	t	P>|t|	95% Conf. Interval	
Volume	1,63E-08	2,53E-08	0,65	0,518	-3,34E-08	6,60E-08
MarketCap	-1,55E-12	1,19E-12	-1,30	0,194	-3,89E-12	7,91E-13
HeldByInsiders	7,476402	1,360252	5,50	0,000	4,801512	10,15129
HeldByInstitutions	7,269081	0,9186325	7,91	0,000	5,462621	9,075541
Constant	-2,773643	0,8689534	-3,19	0,002	-4,482411	-1,064875
					Obs	371
					R-squared	0,1657

To compare the results from the MLR model with the results from the SLR models, we have made the following table in Excel to get a clearer overview of the comparison:

	Coef.	Std. Err	t	P>|t|	95% Conf. Interval		R-squared
Volume, SLR	-5,06E-08	1,94E-08	-2,88	0,004	-9,42E-08	-1,77E-08	0,0219
Constant, SLR	3,992	0,1463	27,28	0	3,7047	4,2802	
Volume, MLR	1,63E-08	2,53E-08	0,65	0,518	-3,34E-08	6,60E-08	0,1657
MarketCap, SLR	-2,28E-12	9,61E-13	-2,37	0,018	-4,17E-12	-3,88E-13	0,015
Constant, SLR	3,933	0,1429	27,52	0	3,6517	4,2138	
MarketCap, MLR	-1,55E-12	1,19E-12	-1,3	0,194	-3,89E-12	7,91E-13	0,1657
HeldByInsiders, SLR	0,2309	1,0685	0,22	0,829	-1,8703	2,3321	0,0001
Constant, SLR	3,8498	0,1646	23,38	0	3,526	4,1636	
HeldByInsiders, MLR	7,476402	1,360252	5,5	0	4,801512	10,15129	0,1657
HeldByInstitutions, SLR	3,9671	0,6726	5,9	0	2,6446	5,2898	0,0862
Constant, SLR	0,5615	0,5767	0,97	0,331	-0,5724	1,6955	
HeldByInstitutions, MLR	7,269081	0,9186325	7,91	0	5,462621	9,075541	0,1657
Constant, MLR	-2,773643	0,8689534	-3,19	0,002	-4,482411	-1,064875	0,1657

This table can also be found in the Appendix, as table 3.1. As we can see from the table, we have some major differences on some of the results in the models. For instance, the volume coefficient has changed from being negative in the SLR model, to being positive in the MLR. Why these differences occur is something we want to look closer at and discuss in the analysis.

As seen from the table, the coefficients for the three variables Volume (only in MLR), HeldByInsiders and HeldByInstitutions, are defined as being negative in our hypothesis, but are positive in the results. Therefore, we have chosen to change our hypothesis for these variables from a negative effect to a positive effect, so that if we reject the null hypothesis we can assume that the results from the model are correct.

4.4 Alternative MLR with logarithmic variables

As we can see from the MLR model in chapter 4.3, we have extremely small coefficients for volume and market cap. These coefficients tell us how much the short ratio will change, with a one-unit change in the independent variable. To avoid such small numbers, which can be hard to interpret at times, we are now going to run the regression model again with these two variables on logarithmic form. By doing this, we get coefficients that tell us how many percentage points the short ratio will change, with a one percentage point change in the independent variable. The model is presented below:

ShortRatio	Coef.	Std. Err	t	P>|t|	95% Conf. Interval	
lnVolume	0,0145578	0,1090602	0,13	0,894	-0,1999054	0,2290211
lnMarketCap	-0,3186505	0,0991263	-3,21	0,001	-0,513579	-0,1237219
HeldByInsiders	6,930523	1,397904	4,96	0,000	4,181592	9,679454
HeldByInstitutions	7,359802	0,8848768	8,32	0,000	5,619721	9,099882
Constant	4,140174	1,996336	2,07	0,039	0,2144455	8,065902
					Obs	371
					R-squared	0,1951

If we compare this MLR model with the previous one in chapter 4.3, we can see that this has a higher R^2 of, 0,1951 compared to the 0,1657 in the other model. All coefficients still have the same sign but have experienced some small changes in values. We will come back to a further interpretation of the model in chapter 5.1. When we do our analysis, we will mainly focus on the MLR model in chapter 4.3, which is our main model. The model including the logarithmic variables is just an alternative model which we could have used to study the percentage change instead of unit change in the volume and market cap variables.

5 Analysis

We now want to interpret the results found in chapter 4 when we ran the regression models. We are going to look at what the results actually tell us, and how we can use these results to answer our research question. When we look at the significance of the results for each variable, we are going to be using a 0.05 significance level to determine whether the results are significant or not, which is a typical significance level to choose. With this significance level, we have a 5% chance to conclude that we falsely reject our null hypothesis of the variable having an effect, when the variable actually has an effect. Thereafter, we are going to be running some tests, a t-test, f-test and test the variables correlation to be more precise, to determine if the variables actually have an effect on short ratio or not. With these tests, we will be able to either reject or fail to reject our null hypothesis that we presented in chapter 1.4. And for the final part, we are going to explain why we got the results that we did, and whether the SLR and MLR assumptions holds in our model.

5.1 Interpretation of the models

We now want to interpret the coefficients from both our SLR and MLR model and discuss its significance level. We want to look at what these coefficients tell us, and how the p-value compared to our significance level can determine whether we reject or fail to reject our null hypothesis from chapter 1.4.

In chapter 3.2, we discussed all the variables, including the Short Ratio. We found that the formula is $\frac{Shares\ Shorted}{Volume}$. If we multiply this with 100, we will get the Short Ratio in percentage of Volume.

To start off, we want to look at the Volume variable. Regarding SLR, one stock increase in volume is estimated to decrease Short Ratio by 0,0000000506. Regarding MLR, one stock increase in volume is estimated to increase Short Ratio by 0,0000000163, with all other variables held equal. As mentioned in chapter 4.2, these numbers may seem small. But, we need to note that changes of only one or two stocks in volume for companies that have an average daily trading volume of millions is like a drop in the ocean, and we do not expect this to have an impact. What is noteworthy on the other hand, is if we look at a change of 10 000 000 stocks in volume. Then the SLR model estimates the short ratio to increase by 0,506 and the MLR model estimates the short ratio to increase by 0,163. We can now see that volume may have a significant impact, even though the coefficient seems low. We find this the most interesting variable to study, because the effect changes direction from the SLR model to the MLR model. For the SLR model, Volume has a significance level (p-value) of 0,004, which is less than our chosen significance level of 0,05. This means that we can reject our null hypothesis at a 5% significance level, because we only have a 0,4% chance to reject a true null hypothesis. In the MLR-model, Volume has a significance level of 0,518, which is much greater than 0,05. This means that in the SLR model, the p-value is significant enough to reject H_0. If we evaluate the MLR, the p-value is not at all significant enough to not reject H_0. We interpret this as if you chose to reject the null hypothesis, in 51,8% of the cases you will reject a true null hypothesis, which ends up being over 50 of 100 times. If we go back to chapter 1.4 with our hypothesis, we remember that our H_0 is that the variable has no effect on the short ratio. Since we cannot reject this hypothesis, there is a fair chance that volume has no effect on the short ratio, according to the MLR model where we have controlled for more variables.

Another interesting observation is that the coefficient changes from being negative in the SLR to being positive in the MLR. If we go back to 3.2, we remember that Volume mathematically is a part of the Short Ratio formula. By being the denominator, it should be mathematically impossible that Short Ratio increases, as a result of an increase in Volume. Even though according to our MLR model, it looks like this is the case. To make sense of this, the numerator (Shares Shorted) also has to increase with relatively more than the denominator. We will come back to an explanation of this later in the analysis.

For market cap, regarding the SLR model, a one dollar increase in market cap is estimated to decrease Short Ratio by 0,00000000000228. Regarding the MLR model, a one dollar increase in market cap is estimated to increase Short Ratio by 0,00000000000155, with all else variables held equal. Also for this variable, the coefficients may seem low. Regardless, if we think intuitively on it, a one dollar change in market cap is not something you may expect be significant to do any impact. Considering some of the companies have several millions stocks and the share price experience oscillations, we can expect from the market cap formula that the market cap also will experience oscillations. To study a significant impact, we can imagine that we have an increase of 1 billion dollars in market cap ($1 000 000 000). The SLR model then estimates the short ratio to increase by 0,00228, and the MLR model estimates the short ratio to increase by 0,00155. We can see that these impacts are still not huge, but for multi-billion, or even trillion-dollar companies, the impact may be significant enough to be noteworthy. If we evaluate the SLR model, the p-value is 0,018, which is less than 0,05, and for the MLR model it is estimated to be 0,194, which is greater than 0,05. So, Market Cap has a significant effect in the SLR model, but not in the MLR model because of our significance level requirements. The coefficients are both positive, regardless of which model we look at, but are bigger in SLR than in MLR. The reason is most likely because of correlative effects with the other independent variables that is unobserved in the SLR model, and Market Cap is therefore not a significant variable in MLR.

For the percentage of shares held by insiders, regarding the SLR model, a one percentage point increase in PSH-insiders is estimated to increase Short Ratio by 0,2309. Regarding the original MLR model, a one percentage point increase in PSH-insiders is estimated to increase Short Ratio by 7,476402, all else variables held equal. In the SLR model, the p-value is estimated to be 0,829, which is much greater than 0,05 and the variable is therefore not significant at all. If we go to the MLR model, the p-value is 0,000, which is much less than 0,05. This means these variables go from being insignificant in the SLR model to being highly significant in MLR. The effect of PSH-insiders is a lot bigger in MLR than SLR, as we can see from the coefficient increasing drastically. That is a sign of SLR being biased, where something unobserved affects PSH-insiders and makes the coefficient smaller than it should be. In the alternative model which includes the logarithmic variables, the coefficient changes to 6,930523. The difference is not significant enough to change our interpretation of the impact, and the variable is still significant according to the p-value.

For the percentage of shares held by institutions, regarding the SLR model, a one percentage point increase in PSH-institutions is estimated to increase short ratio by 3,9671. Regarding the MLR model, a one percentage point increase in PSH-institutions is estimated to increase Short Ratio by 7,269081, all else held equal. The p-value in both SLR and MLR equals 0,000, which is far less than 0,05, and means the coefficient is statistically significant in both models. Like in PSH-insiders, we see an increase in the coefficient when we are using the MLR-model. The difference here is that the PSH-institutions coefficient also is significant in the SLR model. Regarding the coefficient in the SLR, it could still be biased due to the value being so much lower than in the MLR model, but the scope is smaller than for PSH-insiders. For the alternative model including the logarithmic variables, the coefficient changes to 7,359802. The difference here is also not significant enough to change our interpretation of the impact, and the variable is still significant according to the p-value.

As seen from this, the interpretations show us a huge difference in using SLR and MLR. The next step of our analysis will be to discuss why these differences occur and how we can use these results.

5.2 Assumptions evaluation

We now want to evaluate whether the assumptions for the SLR and MLR model that we introduced in chapter 3.1, holds or not. This can have an effect on whether the model is biased or unbiased, which can affect the results we got from the models. We already know that the SLR model has some biases that affect the results, due to other variables correlating with the independent variable. Also, the fact that three of our independent variables actually are endogenous given, makes the MLR model vulnerable for biases. Volume is the only exogenous variable used but has a high correlation with Shares Outstanding, which is included when calculating the other three independent variables. We will come back to these issues later, but they are important to keep in mind when evaluating the model assumptions.

As we remember from chapter 3.1, SLR. 1 to SLR. 5 is the same as MLR. 1 to MLR. 5. We will therefore refer to these five assumptions as MLR. 1 to MLR. 5, to save ourselves some trouble by continuously mentioning both models. We can therefore start with evaluating these assumptions for both the SLR models and the MLR model, and then finally evaluate MLR. 6 for the MLR model.

We can start off with MLR. 1 to MLR. 3. These are the easiest assumptions to evaluate. MLR. 1 says there must be linearity in the model. For both the SLR model and MLR model, this assumption holds, since we have defined the sample regression function in a linear form. MLR. 3 says there must be enough variation in x. We also know that this assumption holds for both models, otherwise Stata would not be able to run the regression analysis. When it comes to MLR. 2, it says we need to have a random sampling, and for this assumption we need to explain a little more in depth. On behalf of the technology sector, we do indeed have a random sampling with a representative selection of observation. We collected a dataset from Yahoo! Finance which included all possible technology firms with data available, listed at the NASDAQ Composite Index. If we narrow our research question down to only study the independent variables effect on short ratio within the technology sector, this assumption holds. But, on the other hand, if we want to look at the effect in the stock market on a general basis, this assumption does not necessary hold, since the technology industry has its own features that other industries might not have. We therefore only represent this sector, and not the market on a general basis. So, to summarize, we can say that MLR. 1 to MLR. 3 holds with this assumption of only representing the technology sector.

For MLR. 4 which states that there should be no correlation between the independent variable and the error term in the model, we already know that this assumption does not hold for the SLR model. Because three of the variables are endogenously given and calculated using shares outstanding, we know they have a correlation with each other. When we only include one of the variables in the model, the others will include in the error term and therefore we have a correlation with the independent variable and the error term. For the MLR model, we have released some of the correlation with the error term, by including these variables in the model. Regardless, we still have more variables that is left out of the model, which is also correlated with the variables in the model. This means we still have some correlation between x and u, but to a smaller extent than in the SLR model. An important aspect to remember is that in practice, we will almost never have a perfect zero correlation. This is due to a phenomenon called omitted-variable bias, which means that variables outside the model have an impact on the variable used in the model. These variables will be included in the error term as long as we do not include them in the model, and therefore we will have a correlation which breaks MLR. 4. Hence, the more variables we include in the model, the more we control for our independent variable, and the more we remove correlations between the independent variable and the error term. We will get back to the omitted-variable bias in chapter 5.4 and explain it more in depth.

MLR. 5 is an assumption of homoscedasticity, which means that the variance should stay the same even though the independent variable increases. This assumption can be tested by running a Breusch-Pagan test in Stata, by using the command “hettest”. The test can be found in table 4.2 in the Appendix. The test is basically a hypothesis test with homoscedasticity as H_0 and heteroscedasticity as H_A. When we ran the test, we got a p-vaule of 0,0001, which tells us that if we reject H_0, there is a 0,01% chance that we reject a true H_0. We can therefore conclude from the test that we can reject H_0, and our MLR model is heteroscedastic. MLR. 5 does therefore not hold, but this does not affect the coefficients. Thus, this can make the p-value, standard deviation and confidence interval be biased.

MLR. 6 is only relevant for the MLR model, and not the SLR. If this assumption holds, MLR. 4 and MLR. 5 will also hold automatically. This assumption states that the error term should be normally distributed with a mean of zero and variance equal to sigma squared. As we concluded, MLR. 4 holds to a certain degree, but not completely. Since we already know that neither MLR. 4 or MLR. 5 holds, we can assume that this assumption does not hold either. We can test this assumption by predicting the residuals in Stata and make a histogram which illustrates the residuals distribution. This histogram can be found in the Appendix in table 6.1, and from it we can see that the distribution is right skewed and not normal. Even though MLR. 4, MLR. 5 and MLR. 6 does not hold, we can assume that our results still are correct enough to interpret and analyse, even though they might be biased.

5.3 Analytical discussion and testing

Finally, we can now proceed to analyse what these results actually mean. As mentioned in 5.1, the results from the four different SLR-regressions and the common MLR-regression gives completely different results. By using different extra tests, such as f-tests and correlation tests, the hope is to clear up the obscure picture that the regression models give us. The 6 MLR assumptions have already given some answers.

Starting off we want to look at the correlation tests which can be found in table 4.1 in the Appendix. The correlation between the independent variables is tested, and we also included the variable “shares outstanding” because it is included in the mathematical formula of our independent variables. In this test, the most interesting result is shares outstanding. Shares outstanding is the total amount of stocks for the firm. As explained in 3.2, these variables are given in three variables that hold shares outstanding. The correlation is measured as a correlation score, that goes from -1 to 1. If we have -1 it means perfectly negative correlation, while 1 means perfectly positive correlation. 0 means no correlation at all. A high correlation (close to -1 or 1) is not necessarily good or bad, it depends on the goal for the research. High correlation could explain some of the biases that it's discovered in SLR compared to MLR.

As seen in table 4.1 in the Appendix, one can see the correlation between those four independent variables. It shows that market cap and volume have a correlation on 0,66. PSH-insiders and PSH-institutions also have 0,66. Those are the two variables with the highest correlation effect, of the independent variables. None of them is extremely high, but it is still noteworthy. The correlation between the two PSH variables and Volume is not extraordinary. By taking shares outstanding into account, the picture changes. Shares outstanding has a remarkable positive correlation with Volume (0,78) and Market Cap (0,92). For market cap, there is almost a perfect correlation with shares outstanding. Not a very surprising result, given that shares outstanding is a part of the market cap formula, as seen in chapter 3.2. In our alternative mode, where we have used the logarithmic variables for volume and market cap, the correlation with shares outstanding is weaker. Approximately 0,36 for lnVolume and 0,4 for lnMarketCap. This correlation test can be found in table 4.3 in the Appendix.

More interesting is the high correlation between volume and shares outstanding, at almost 0,8. It is not a perfect correlation, but it is that high so it can explain the differences found in the models. By running a SLR model of the shares outstanding’s effect on volume, the regression model shown in table 1.5 in the Appendix tells us that one stock increase in shares outstanding will increase volume by 0,0055. This is a significant result with the adjusted R2at 0,616, which means that shares outstanding explains 61,6% of the variation in the volume variable. This SLR of course has its biases itself, but it makes it possible to understand the violations in volume from SLR and MLR. It makes a strong suspicion that MLR assumption

4, $E(u|x) = 0$ does not hold. By looking at the correlation test together with this regression, it is enough evidence to write volume as a standard sample regression function of shares outstanding.

Volume = $\beta_0 + \beta_1$SharesOutstanding + u

From part 3.2, we saw that the three other independent variables all are equations dependent on shares outstanding. By rewriting these three equations and the new last equation for volume, as solved for share outstanding, we get:

From market cap, => Shares outstanding = $\frac{Market\ cap}{Stock\ price}$

From PSH-Insiders, => Shares outstanding = $\frac{Shares\ held\ by\ insiders}{Percentage\ of\ shares\ held\ by\ insiders}$

From PSH-Institutions, => Shares outstanding = $\frac{Shares\ held\ by\ institutions}{Percentage\ of\ shares\ held\ by\ institutions}$

From volume, => Shares outstanding = $\frac{Volume - \beta_0 - u}{\beta_1}$

The reason these equations are presented is to show that all these variables can be written endogenous as functions of each other, and makes us believe that the unobserved variables in the SLR models correlate with x.

The F-test tells about the significance of the joint effect of the variables. Since there are 4 independent variables, it is possible to test two or three variables together to see if their joint effects are significant on the dependent variable. Even if all of the independent variables were insignificant independently, they could have a significant effect on y together. The total f-value of the MLR models seems to be significant. So, these four independent variables together have a significant effect on Short Ratio. Volume was not significant in MLR alone. Together with PSH-insiders and PSH-institutions, it will have a significant effect with a p-value at 0,000. While running f-tests together with market cap, the market cap variable will be dropped by Stata. The reason for this might be the extremely high correlation between market cap and shares outstanding.

5.4 Omitted-variable bias

A way to explain the results found in the volume variable where the coefficient changes from being negative in the SLR model to being positive in the MLR model, is due to a phenomenon called *Omitted-variable bias*. This occurs when a statistical model leaves out one or more relevant variables. The bias results in the model attributing the effect of the missing variables to those that were included (*Omitted-variable bias*, 2020). When we ran the MLR regressions, we tried every possible combination with the independent variables on the short ratio variable. This resulted in a total of 11 different MLR models, and 4 different SLR models, which all can be found in the Appendix. What we found then, is that for all the combinations that included volume, the coefficient for volume changed drastically from the SLR model, but only in the model including all four variables was it positive. This means that for all the MLR models where we left out at least one variable, the coefficient was still negative like in the SLR model.

So, to explain this in with a practical example, imagine you go for a run. You have a pulse watch which tracks information about your pulse, your speed and the grounds' incline. When you get home, you want to try and run a regression analysis to explain your pulses behaviour. You start off by running an SLR model including only speed, and finds that there is a negative relationship between your pulse and speed. Intuitively, this does not make sense, because one would assume that the pulse increases when you run faster. So, you run another regression analysis, this time including incline as well which makes the model a MLR model. You now see that the coefficient for speed has changed and become positive, while there is a positive relationship between incline and pulse as well. What has happened here is that in practise when you run, your pulse will increase drastically when running uphill, while your speed may reduce a bit at the same time. When running the SLR model, the model does not capture the incline and therefore estimates there to be a negative relationship between speed and pulse, due to there being a negative relationship between incline and speed. When including incline in the model, this variable is controlled for and does not affect the speed variable like it did in the SLR model. Therefore, we will now get a more correct coefficient for speed which has a positive relationship with pulse.

Essentially, this is what we believe has happened to volume in our model. In the SLR model, there are other factors that affect volume and we therefore get an incorrect relationship between volume and short ratio. The interesting part on the other hand, is that we actually intuitively expect volume to have a negative relationship with short ratio. From the formula for short ratio, we found that:

Short ratio = $\frac{shares\ shorted}{volume}$

If volume increases, it mathematically follows that the short ratio should decrease. In the MLR model, we instead find that there is a positive relationship between the two variables. So, at first glance, it looks like the SLR model has a more correct coefficient than the MLR model, which is the opposite result as with the omitted-variable bias scenario we explained above. What is important to notice though, is that the p-value tells us that the result in the SLR model is significant, while in the MLR model it is insignificant. We believe the reason for this is that the independent variables that we have used in the MLR model, are all calculated using shares outstanding as a variable. When running a correlation test, we found that volume is highly correlated with shares outstanding, and we therefore get this change in effect. The correlation test can be found in table 4.1 in the Appendix, and we found that volume and shares outstanding are 0,7848 correlated, which is quite close to 1 (perfect correlation). The alternative model might not have this problem, because the logarithmic variable does not have the same high correlation with shares outstanding. That means the coefficients are more independent from this “endogenous problem”.

6 Conclusion and summary

In this assignment, we have tried to determine four different variables' effect on the short ratio. First using a SLR model for all the variables individually, then using a MLR model including all four variables, and at last using an alternative MLR model where we changed two of the variables to their logarithmic form. We have found interesting and significant results, with evidence enough to say that the short ratio is not just a random variable, but can be explained by other financial variables. The four variables market cap, volume, % of shares held by insiders and % of shares held by institutions, have different effects and not all of them are significant in itself, but all together have a significant effect on short ratio. The MLR model with these four variables had a R-squared of 0,1657, which means that these four variables explain 16,57% of the short ratio variable. If we had included more variables in the model, we could assume that we would have gotten a larger R-squared. Still, even by including all variables that can have an effect on the short ratio, we would probably not have reached an R-squared of 1. This is due to the fact that some variables cannot be measured with numbers, but still have an effect on the short ratio. We will come back to an example of such a variable.

The inspiration to write about this specific research question was the GameStop-incident. GameStop had a high short ratio when the incident took place. This might be because one of the four variables we studied influenced it, or because of a joint effect combined by multiple variables. GameStop-shorters could have been suffering because GameStop had a relatively low market cap, or that PSH-institutions or PSH-insiders was high. For these variables, we are sure about the coefficient and in which direction it affects the short ratio. It's still important to understand that the incident had other factors that were determining, but that cannot be quantified inside the financial market. Like the David vs Goliath-reference we mentioned in the introduction. Small investors might be more inspired to contribute to make a short squeeze than they normally would, just to create losses for the big investors. This inspiration and motivation cannot be included in our MLR model, but still could affect how large the short ratio will be. This is one of the variables mentioned above that we cannot measure with numbers, and is the reason the R-squared cannot be equal to 1.

In chapter 2.1, we presented three interesting papers related to our research question. Cohen, Diether and Malloy conclude that neither volume nor institutional ownership have any significant abnormal effect of short return. However, in our analysis we find that institutional ownership has a significant effect on the short ratio. These results do not need to be contradictory, since we do not know the effect of short ratio on the return. Itamar Drechsler and Qingyi Freda Drechsler found that a large and statistically significant effect from shorting on the average return. So, it seems that the findings in the two papers were two opposite effects, and the results may be determined by which variables and approach they used in their papers. Dr. Usman Chohans paper about the GameStop incident, is a paper we used in chapter 2.4 about the GameStop incident to give us a better overview and explanation of the incident.

Our model's estimates alone are not enough to be able to predict what the short ratio will be. People have tried to predict the financial markets with varying success. Regardless, our findings may help understand in which direction a certain shock in one of the independent variables we used, can be expected to affect the short ratio. To make our estimates more precise and the model less biased, we could have included several other variables. However, this would have taken a lot more time to pull through, and to stay within our timeframe and restrictions for this assignment, we chose to not include more than four.

In this assignment, we have not discussed the ethics of shorting. Is it ethical to speculate in the financial market? To yield money at companies going bad, potentially bankruptcy, and workers get unemployed, and so on. The GameStop incident might have been an eyeopener in terms of consequences of shorting and short-squeezes. By writing this assignment, we could help to enlighten these sides of shorting. We have also enlightened how the short ratio can be predicted by studying other financial parameters. While we do not have any conclusions or recommendable measures on how to avoid short squeezes, we can recommend to at least be careful and aware of the risk that comes with shorting companies with a high short ratio. That again can be found by looking at market cap, shares held by insiders and shares held by institutions, while the volume is not a good variable for an evaluation like this.

Bibliography

- Chen, J (2021) *Short Interest Ratio*
  Availible at: https://www.investopedia.com/terms/s/shortinterestratio.asp
  [Accessed 23rd of March 2021]

- Chohan, U.W. (2021) *Counter-Hegemonic Finance: The GameStop Short Squeeze*
  Doi: http://dx.doi.org/10.2139/ssrn.3775127
  [Accessed 28th of April 2021]

- Cohen, L., Diether, K.B. and Malloy, C.J. (2007). *Supply and demand shifts in the shorting market*
  Doi: https://doi.org/10.1111/j.1540-6261.2007.01269.x
  [Accessed 28th of April 2021]

- Drechsler, I. and Drechsler, Q.F. (2014). *The shorting premium and asset pricing anomalies*
  Doi: 10.3386/w20282
  [Accessed 28th of April 2021]

- Duggan, W. (2019) *12 Stocks with short ratios above 50*
  Availible at: https://finance.yahoo.com/news/12-stocks-short-ratios-above-150409195.html
  [Accessed 22nd of April 2021]

- Fernando, J. and Boyle, M. (2021) *Market Capitalization*
  Availible at: https://www.investopedia.com/terms/m/marketcapitalization.asp
  [Accessed 7th of April 2021]

- *F-test* (2021)
  Availible at: https://en.wikipedia.org/wiki/F-test
  [Accessed 16th of April 2021]

- Grant, K. (2021) *GameStop: What is it and is it trending?*
  Availible at: https://www.bbc.com/news/newsbeat-55841719
  [Accessed 22nd of April 2021]

- Hayes, A. and Anderson, S. (2021) *Volume*
  Availible at: https://www.investopedia.com/terms/v/volume.asp
  [Accessed 3rd of March 2021]

- Jagerson, J. (2017) *Understanding Institutional Ownership*
  Availible at: https://www.learningmarkets.com/understanding-institutional-ownership/
  [Accessed 21st of April 2021]

- Kenton, W. (2020) *Institutional Ownership*
  Availible at: https://www.investopedia.com/terms/i/institutional-ownership.asp
  [Accessed 21st of April 2021]

- Kolhatkar, S. (2021) *What happened with GameStop stock prices? Why did their stock prices rocket up?*
  Availible at: https://www.republicworld.com/technology-news/gaming/what-happened-with-gamestop-stock-prices-why-did-their-stock-prices-rocket-up.html
  [Accessed 22nd of April 2021]

- *Linear regression* (2021)
  Availible at: https://en.wikipedia.org/wiki/Linear_regression
  [Accessed 26th of March 2021]

- Loth, R. (2019) *How can institutional holdings be more than 100%?*
  Availible at: https://www.investopedia.com/ask/answers/07/institutional_holdings.asp
  [Accessed 21st of April 2021]

- Mitchell, C. and Scott, G. (2021) *Short squeeze*
  Availible at: https://www.investopedia.com/terms/s/shortsqueeze.asp
  [Accessed 23rd of March 2021]

- Musk, E. (2021) Twitter
  Availible at: https://twitter.com/elonmusk/status/1354890601649610753
  [Accessed 10th of February 2021]

- Nordnet (2020) *Shorthandel*
  Availible at: https://www.nordnet.no/blogg/borsskolen/laer-om-aksjer/shorte-aksjer/
  [Accessed 23rd of March 2021]

- *Omitted-variable bias* (2020)
  Availible at: https://en.wikipedia.org/wiki/Omitted-variable_bias
  [Accessed 29th of April 2021]

- *Ordinary least squares* (2021)
  Availible at: https://en.wikipedia.org/wiki/Ordinary_least_squares
  [Accessed 26th of March 2021]

- *P-value* (2021)
  Availible at: https://en.wikipedia.org/wiki/P-value
  [Accessed 16th of April 2021]

- *Shares Owned By Insiders* (2021)
  Availible at: https://ycharts.com/glossary/terms/shares_owned_by_insiders
  [Accessed 21st of April 2021]

- *Short (finance)* (2021)
  Availible at: https://en.wikipedia.org/wiki/Short_(finance)
  [Accessed 9th of March 2021]

- Statista (2021) *Weekly NASDAQ Composite index performance 2021*
  Availible at: https://www.statista.com/statistics/1104283/weekly-nasdaq-index-performance/
  [Accessed 1st of March 2021]

- *T-test* (2018)
  Availible at: https://no.wikipedia.org/wiki/T-test
  [Accessed 16th of April 2021]

- *Why use short ETPs to achieve a short position?* (2017)
  Availible at: https://www.wisdomtree.eu/en-gb/-/media/eu-media-files/other-documents/educational/inverse-etps--v-short-selling-final-(eng)-sep2017.pdf
  [Accessed 23rd of March 2021]