The Impact of Macroeconomic Variables on Stock Return in Different Industries - A Multiple Linear Regression

ELIN PALMGREN

NATASHA NANAKORN
The Impact of Macroeconomic Variables on Stock Return in Different Industries

- A Multiple Linear Regression

ELIN PALMGREN

NATASHA NANAKORN

Degree Projects in Applied Mathematics and Industrial Economics (15 hp)
Degree Programme in Industrial Engineering and Management (300 hp)
KTH Royal Institute of Technology year 2019
Supervisors at KTH: Tatjana Pavlenko och Julia Liljegren
Examiner at KTH: Jörgen Säve-Söderbergh
Abstract

Macroeconomics constitute a central part of fundamental analysis of stock markets and consequently the relationship between macroeconomic variables and stock markets is far from questioned. However, there is no general consensus regarding neither the extent of this relationship nor whether the relationship varies amongst industries. The aim of this thesis is therefore to determine the macroeconomic variables most important in explaining variations in stock return within two separate industries and furthermore the share of these variations solely accounted for by macroeconomic variables. To this end, a multiple linear regression approach is used and Nasdaq indexes OMX Stockholm Industrial Goods & Services and OMX Stockholm Banks are used as proxies for the two selected industries.

The final result of this analysis is that the variables repo rate, SEK/EUR exchange rate, consumer expectations, oil price, GDP, money supply and inflation are statistically significant in explaining stock return within industrial goods and services whilst SEK/USD exchange rate, SEK/EUR exchange rate, oil price, GDP, money supply and inflation are statistically significant in explaining stock return within the banking industry. The analysis of the extent of the impact of these variables on stock return is, however, deemed inconclusive due to time dependencies amongst the variables.
Sammanfattning


Det slutgiltiga resultatet av denna analys är att variablerna reporänta, växelkurs SEK/EUR, konsumentförväntningar, oljepris, BNP, penningmängd och inflation är statistiskt signifika för aktieavkastning inom industrin, medan växelkurs SEK/EUR samt SEK/USD, oljepris, BNP, penningmängd och inflation är statistiskt signifika för aktieavkastning inom bankbranschen. Analysen av omfattningen av dessa makrovariablers påverkan på aktieavkastning när emellertid ingen slutsats, vilket anses kunna hänföras till tidsberoende hos variablerna.
Acknowledgements

We would like to express our sincere gratitude to our supervisors at the Royal Institute of Technology, Tatjana Pavlenko at the Department of Mathematics and Julia Liljegren at the Department of Industrial Economics and Management, for their invaluable guidance and support throughout the process of writing this bachelor thesis.


Contents

1 Introduction ................................................. 6
   1.1 Background ........................................... 6
   1.2 Problem Statement ................................... 6
   1.3 Purpose ............................................... 6
   1.4 Scope ............................................... 6

2 Economic Theory ........................................... 7
   2.1 Variables .............................................. 7
      2.1.1 Response Variables ............................... 7
      2.1.2 Regressor Variables ............................. 7
         2.1.2.1 Gross Domestic Product .................... 7
         2.1.2.2 Money Supply ................................. 8
         2.1.2.3 Repo Rate .................................. 8
         2.1.2.4 Inflation .................................... 9
         2.1.2.5 Unemployment Rate ......................... 9
         2.1.2.6 Consumer Expectations ..................... 9
         2.1.2.7 Spot Price of Oil ......................... 10
         2.1.2.8 Exchange Rates ............................. 10
   2.2 Literature Review ..................................... 10

3 Mathematical Theory ...................................... 12
   3.1 Multiple Linear Regression ......................... 12
      3.1.1 Model ........................................... 12
      3.1.2 Assumptions .................................... 13
      3.1.3 Ordinary Least Squares ....................... 13
   3.2 Model Adequacy ....................................... 14
      3.2.1 Residual Rescaling ............................. 14
      3.2.2 Residual Plots ................................ 14
      3.2.3 Leverage and Influence ....................... 15
         3.2.3.1 Measure of Leverage ..................... 15
         3.2.3.2 Measure of Influence .................... 16
         3.2.3.3 Measure of Model Performance .......... 16
   3.3 Model Properties ..................................... 17
      3.3.1 Multicollinearity ............................... 17
      3.3.2 Endogeneity .................................... 18
      3.3.3 Heteroscedasticity ............................ 18
      3.3.4 Analysis of Variance ......................... 18
      3.3.5 Coefficient of Determination ............... 18
      3.3.6 Hypothesis Testing ............................ 19
         3.3.6.1 F-Statistic ................................ 19
         3.3.6.2 t-Statistic ................................. 19
      3.3.7 Confidence Intervals ......................... 19
   3.4 Biased Regression ................................... 20
   3.5 Variable Selection and Model Building ............... 20
      3.5.1 Method .......................................... 20
      3.5.2 Criteria ........................................ 21
1 Introduction

1.1 Background
Each month Swedish investors and analysts anxiously await the release of key macroeconomic indicators such as consumer price indexes and unemployment reports as well as the Swedish central bank’s decision on whether to increase or decrease the repo rate. These indicators jointly form a central part of the fundamental analysis of stock markets, given that the performance of listed companies is very much dependent on macroeconomic conditions for growth. Even though the impact of macroeconomic variables on stock return is far from questioned, there is no general consensus regarding the extent of this impact. [18]

1.2 Problem Statement
The subject of this thesis is to investigate and quantify the impact of macroeconomic variables such as GDP, inflation, exchange rates, unemployment and the repo rate on stock returns on the Stockholm Stock Exchange using multiple linear regression. Furthermore, it is acknowledged that the nature and sensitivity to macroeconomic factors may reasonably vary between industries.[18] For this reason, this thesis will focus solely on stocks within two selected industries rather than stocks listed on the Stockholm Stock Exchange in general, with the hypothesis that the results will vary between the industries. The chosen industries two of the largest industries in Sweden - the manufacturing industry and the banking industry.

The research questions to be answered are consequently as follows:

- Which macroeconomic variables are important in explaining stock return fluctuations within banking and industrial goods and services?
- To what extent do macroeconomic variables explain stock return fluctuations within banking and industrial goods and services?

1.3 Purpose
The aim of the thesis is to determine the share of stock return fluctuations that can be explained solely by macroeconomic variables, independent of any industry or company specific information, and conclude which variables are in fact influential with regard to stock return within different industries. The potential impact of the thesis is increased by the fact that most economic analysts agree that the economy is currently approaching the next recession. As a consequence, everyone that is exposed to the Swedish stock market in one way or another, among them approximately one fifth of the Swedish population [10], has to reconsider the extent and risk of this exposure. Consequently, a quantification of the impact of macroeconomic variables on stock return within different industries may well benefit investors in this decision.

1.4 Scope
To begin with, this thesis is limited to focusing only on stocks listed on the Stockholm Stock Exchange. This limitation is motivated by the fact that most macroeconomic factors such
as GDP, inflation and unemployment rates are country specific and thus mainly impact the local stock exchange.

Another delimitation of the thesis is to focus solely on two selected industries – industrial goods and services and banks. The purpose of this delimitation is primarily to specify the scope of the thesis and thus further the prospect of drawing impactful conclusions, whilst still allowing for a comparison. Including additional industries would naturally benefit the comparison aspect of the thesis but also result in an analysis characterized by width rather than depth. The industries are selected on the basis of market size and diversification in the purpose of reducing potential bias in the results and are simulated by examining Nasdaq indexes *OMX Stockholm Industrial Goods & Services* and *OMX Stockholm Banks*.

Furthermore, the scope of the thesis is limited to data from the period 2001:M1-2018:M09. The time period is long enough to provide a sufficient number of observations as well as include both economic recessions and expansions.

## 2 Economic Theory

In this section, the selected response and regressor variables subject to this thesis are presented and defined. Furthermore, the theoretical relation to stock markets and stock return will be presented, if existent, for each regressor variable. Subsequently, a literature review of existent research will be presented. Key findings from this review will later on be used as foundation for the discussion.

### 2.1 Variables

#### 2.1.1 Response Variables

Given that two response variables are the subject of this thesis, two independent analyses will be performed. The response variables are closing prices of two indexes representing stocks listed on the Stockholm Stock Exchange within two different industries. As mentioned, the selected indexes are *OMX Stockholm Industrial Goods & Services* and *OMX Stockholm Banks*. In the future, the response variables will be referred to as *industry index* and *bank index* or in general terms as *stock return*.

#### 2.1.2 Regressor Variables

The selected regressor variables are *gross domestic product (GDP)*, *money supply*, *repo rate*, *inflation*, *unemployment rate*, *consumer expectations*, *spot price of oil*, *SEK/EUR exchange rate* and *SEK/USD exchange rate*. The basis on which these variables have been selected is presented in 4. Methodology.

#### 2.1.2.1 Gross Domestic Product

The *Gross Domestic Product (GDP)* is defined as the value of all final goods and services produced during a given time period, normally a year, and in a specific country. As a consequence, it is a fundamental measure of total output, where each component is valued at its market price and accumulated. [7, p. 23] Moreover, a country’s GDP can be decomposed into its main components. For Sweden, the following decomposition holds at least approximately: [5, p. 30]
• Consumption (approx. 50% of GDP)
• Investment (approx. 25% of GDP)
• Governmental purchases (approx. 20% of GDP)
• Net exports (approx. 5% of GDP)

Furthermore, the rate at which the GDP is increasing is normally referred to as the growth rate of the economy in general. There are two main drivers of increases in GDP: the available amount of factors of production, the principal ones being labor force and capital stock, and productivity increases. However, a fundamental assumption within macroeconomics is that factors of production are fully employed in the long-run. The hypothesized path of GDP at full employment is commonly called the trend path of GDP. It is around this trend line that actual output fluctuates according to the business cycle, resulting in expansions and recessions. [7, p. 10-12]

2.1.2.2 Money Supply

The money supply of an economy can be evaluated using different measurements. This thesis treats the money supply defined by the measure M1, which includes the most liquid portions of the money supply. This category involves physical currency in circulation and transaction accounts, which refers to accounts to and from which the owner can make direct transactions.[5, p. 53]

Algebraically, the demand of money, M^d, is defined by the following equation, where P denotes price level, Y denotes real output, i denotes interest rate and L denotes liquidity:

\[ M^d = P \times Y \times L(i) \]  

In words, the theoretical relation, commonly referred to as the LM relation, states that the demand of money is equal to the nominal income, defined as the product of price level and real output, multiplied by a function of the interest rate. This function is formulated as to reflect the negative relationship between the demand of money and the interest rate. Furthermore, in the long run, the money market is in a state of equilibrium, which implies that the demand of money equals the money supplied by the central bank. [5, p. 54-57]

2.1.2.3 Repo Rate

Since 1994, Sweden’s policy rate has been referred to as the repo rate. This rate is defined as the interest rate at which financial intermediaries can deposit or borrow funds at the Swedish central bank (Riksbanken) for a period of seven days. [36] Theoretically, the interest is defined by the state of equilibrium of the money market, that is when the demand of money equals the money supplied by the central bank. [5, p. 56-57]

In practice, the rate is determined by adjusting the money supply. An increase in money supply implies a decrease in interest rate and vice versa. Such adjustments in the money supply are performed by the central bank using open market operations, where the central bank buying bonds is referred to as expansionary monetary policy and selling bonds as contractionary monetary policy. [5, p. 56-61]
Beyond money supply and output, the interest rate is also related to investments, where investments refer solely to those performed by companies. This relation is likewise negative, implying that an increase in interest leads to a decrease in investment and vice versa. In turn, output depends on investment through the IS relation. [5, p. 78-88]

2.1.2.4 Inflation

Inflation is defined as the continuous increase in price level and the rate of this increase is referred to as the inflation rate. There are several methods for measuring the inflation rate but the most common measurement is the Consumer Price Index (CPI), which indicates the average price of the consumption of households and is the measurement used in this thesis. [5, p. 18-19]

If inflation exclusively entailed an even, stable and predictable increase of the general price level in the economy - commonly referred to as pure inflation - then inflation would not be such a complicated and debated phenomenon. In practice, pure inflation does not exist, [5, p. 18-19] which is why one of the main focuses of Riksbanken is to maintain a target inflation of 2%. The motivation provided by the central bank is that such a target contributes to price stability and favorable conditions for long-term sustainable growth, as well as a benchmark for inflation expectations. [35]

2.1.2.5 Unemployment Rate

An unemployed person is someone who is able to work but is currently not employed. Subsequently, unemployment is measured as the share of the labour that is unemployed. Labor, in turn, includes everyone within in the age-span 15-74 that are either unemployed or employed. For instance, pensioners, full-time students or sick people are not unemployed but simply not considered a part of the labor force. Further on, the number of people that are unemployed is affected by the current state of the economy in Sweden as well as that of the global economy. [38] For instance, the well renowned graphical model known as the Phillips curve illustrates that there is a relationship between the rate of unemployment and the rate of inflation. More specifically, the Phillips curve states that there is a trade-off between the rate of unemployment and the rate of inflation; a decrease in the rate of unemployment will correlate with a higher rate of inflation and vice versa. [5, p. 122]

2.1.2.6 Consumer Expectations

Currently there is no general consensus regarding the exact definition of consumer expectations. In this report, the Swedish National Institute of Economic Research’s (Konjunkturinstitutet) economic tendency survey regarding household indicators, more specifically the consumer confidence indicator, will be used as a measurement of consumer expectations. The survey targets Swedish households and contains four questions regarding the development of the Swedish economy, covering subjects such as the general state of the economy, unemployment and price levels. In addition to the four general questions, the survey also contains questions that investigate the willingness to consume. The answers are weighted numerically and the indicators are standardized with a mean of 100 and standard deviation of 10. [26]
2.1.2.7 Spot Price of Oil

Today, the modern economy is very much dependent on non-renewable fossil fuels and one of the most important ones is crude oil. The dependence on these resources is interesting and relevant to this analysis since the supply is influenced by political forces and consequently occasionally volatile. This is because the major suppliers have formed an intergovernmental organization in order to coordinate petroleum policies, leading to major influence on global oil prices that historically have resulted in major supply shocks with severe effects on global economy. [20] One of the greatest supply shocks of all times was one caused by oil during the 1970s. Generally, supply shocks result in both an increased general price level and a decreased level of output. [7, p. 139-142]

However, oil is not only physically utilized within manufacturing and transportation but also a high frequency commodity traded on the stock exchange. For this purpose, a benchmark is used as reference price for buyers and sellers of crude oil. One of the most common such benchmarks is West Texas Intermediate (WTI) crude oil, which is traded on the New York Mercantile Exchange. [21] It is furthermore this reference price that has been used as the spot price of oil in this analysis.

2.1.2.8 Exchange Rates

An exchange rate is defined as the rate at which one currency can be exchanged for another. Furthermore, exchange rates are affected by the international balance of payments. If a country has a large trade deficit during a longer period of time, then eventually the currency will lose its value relative to other currencies. Naturally, the opposite holds for trade surplus. [7, p. 295-297] Since Sweden is a country that to a high extent is dependent on foreign trade it is natural to consider two of the most central currencies, i.e. the USD and EUR currencies. Further on, exchanges rates are traded on the foreign exchange market and the spot exchange rate represents the current exchange rate at a specific time for a specific currency. The data representing the exchange rates used in this thesis are the spot exchange rates taken from Avanza.

2.2 Literature Review

A number of studies examining the relationship between macroeconomic variables and stock return have been performed during the last decades. Even though the findings do not necessarily align when it comes to the nature of the correlation, there is more or less consensus within financial research that stock markets are not unaffected by macroeconomic factors. For example, Muradoglu, Taskin, and Bigan (2000) examined 19 emerging markets over a period of twenty years and concluded that macroeconomic variables such as inflation, exchange rates and interest rates are indeed significant for stock return, even though they also found that the relationship was dependent on the local stock market’s integration with other markets and its relative size. [33] Others who have established some kind of relationship between macroeconomic variables and stock return will be referred to successively throughout this literature study, with more focus on results for specific macroeconomic variables.

To begin with, inflation has been subject to much financial research and study when it comes to its effect on stock markets. Already in the 1970s, Fama and Schwert (1977)
published results indicating a negative relationship between inflation and stock return. [12] Since then, various other publications have established the same relationship on markets ranging from the United States to China and Malaysia (Feldstein (1980)[13], Gallagher and Taylor (2002)[14], Geetha et al (2011)[15]). Spyrou (2004), on the other hand, contradicts previous research and presents a positive relation between inflation and stock return for a number of emerging markets.[41] Graham (1996) performs a more multifaceted analysis and concludes that whether or not the relationship between inflation and U.S. stock return is positive or negative depends on the current monetary policy. [16]

Other macroeconomic factors closely related to monetary policy are money supply and policy rate. For example, Ekanayake, Rance, Halkides (2008) demonstrate a negative relationship between the U.S. policy rate, referred to as the federal funds rate, and stock return. [8] Kontonikas, MacDonald, Saggu (2013) reach the same conclusion for the U.S. market for periods of financial expansion. During financial crisis, on the other hand, the authors find that stock return did not increase as a result of unexpected federal fund rate cuts. [27] Kuttner and Bernanke (2005) furthermore analyzes the impact of unexpected monetary policy actions on stock return using federal funds futures data and confirms the conclusion that there is a negative relationship between federal funds rate and stock return. [4] The federal funds rate is probably one of the most well-known monetary policy tools but another important component is money supply. [7, p. 196-197] Despite this fact, the relationship between money supply and stock return is not as well-explored as that between inflation or federal funds rate and stock return. Fama (1981) and Sorensen (1982) are, however, examples of researchers who have studied this relationship and both demonstrate a direct relationship between money supply and stock return. [11] [40] Rjoub, Tursoy, Gunsel (2009), on the other hand, have examined the effect of macroeconomic factors on thirteen stock portfolios on different markets and found a positive relationship between money supply and stock return in seven of them but a negative relationship in the rest of them. [37]

Another article where monetary policy is the subject of analysis is Bernanke and Kuttner’s (2004). They demonstrate a negative relationship between the federal funds rate and stock return if unemployment reports are not presented simultaneously. If decisions regarding increase or decrease of rate is presented along with the release of the unemployment report, on the other hand, the effect on stock return was close to nonexistent. [4] Boyd, Hu och Jagannathan (2005), on the other hand, concludes that reports of increased unemployment affects stock return positively during periods of economic expansion but negatively during periods of economic recession. [6] Zoega (2012) supports the conclusion of a negative correlation between unemployment and stock return but also emphasizes that the results are closely related to expectations and uncertainty about the future. [43]

Apart from Zoega (2012), a number of others have investigated the relationship between stock return and behavioral psychology. For example, Altunas, Mera, Sarikovanlik (2017) use vector autoregression to investigate psychological factors and their effect on financial markets and find a long-term causality relationship between investor sentiment and stock return. [2] This is relevant since investors are also consumers. Furthermore,伊西奥格鲁 (2012) establishes both a short- and long-term relationship between consumer expectations and stock return on Turkish and Spanish markets. [19]

Furthermore, consumers’ expectations on the economy are closely related to economic growth and GDP. Pece (2015) applied methods such as time series analysis and Granger causality and presented results indicating both that portfolio investments have a positive
impact on GDP growth and that GDP growth has a positive impact on stock return. [34] The long-term positive relationship between GDP and stock return has been confirmed by a number of studies, where two examples are Marozva, Magwedere (2017) for the Johannesburg Stock Exchange [31] and Khan and Zaman (2012) for the Karachi Stock Exchange. [25]

Beyond economic growth, Khan and Zaman (2012) were also able to indicate a positive relationship between exchange rates and stock return. [25] This result is confirmed by Aggarwal (1981), who examined the relationship between a trade-weighted value of the U.S. dollar and the U.S. stock market indexes during the 1970s. [1] In contrast, Soenen and Hennigar (1988), also focusing on the U.S exchange and stock markets, present a negative relationship between a currency-weighted value of the U.S. dollar and U.S. stock market indexes. [39] Muradoglu, Taskin and Bigan (2000), on the other hand, analyzed a number of emerging markets and concluded that exchange rates, among other macroeconomic factors, are positively related with stock return but also highlighted that the relationship was dependent on the stock market’s integration with other markets and its relative size. [33]

Further on, Basher, Haug, Sadorsky (2012) present results indicating that exchange rates are not only related to stock return but also to the spot price of oil. In particular, the results indicate that positive shocks to the spot price of oil have negative impact on both stock return and USD exchange rates, at least in the short run. [3] This negative relationship between oil price and stock return is furthermore supported by results presented by Kang, Ratti, Yoon (2015). [24]

In summary, there is published and more or less established support for the hypothesis that the chosen macroeconomic variables do in fact have an impact on stock return. For most variables – mainly policy rate, unemployment, consumer expectation, GDP and oil price – there is more or less consensus amongst economists regarding the relationship between the macroeconomic variable and the stock return. For others – mainly inflation, money supply and exchange rates – the existing results are less unanimous.

3 Mathematical Theory

In this section, the mathematical theory behind multiple linear regression analysis is presented, necessary for understanding and drawing conclusions from the results and findings of the thesis.

3.1 Multiple Linear Regression

3.1.1 Model

A multiple linear regression model relates a response variable y to k regressor variables, x_j, j = 1, . . . , k, using p = k+1 regression coefficients, \( \beta_j \), j=0,1,...,k. Each \( \beta_j \) describes the change in y per unit change in x_j holding all other regressor variables x_i, i \( \neq \) j, constant. Such a model can be expressed either for each observation i = 1, . . . , n or using complete matrix notation:

\[
y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \quad (2)
\]
\[ y = X\beta + e \] (3)

As stated by above definition, a linear regression model is a regression model that is linear in the regression coefficients, the \( \beta \)'s, but not necessarily in the regression variables, the \( x \)'s. [32, p. 67-70]

### 3.1.2 Assumptions

A somewhat stable regression model requires that five assumptions be made and satisfied. If violated, the consequences may be rather severe in terms of stability and reliability of the model itself. [32, p. 129]

1. There is approximately a linear relationship between the response and the regressors
2. The mean of the errors is zero, i.e. \( E[\varepsilon_i] = 0, i = 1, \ldots, n \)
3. The variance of the errors is constant, i.e. \( \text{Var}[\varepsilon_i] = \sigma^2, i = 1, \ldots, n \)
4. The errors are uncorrelated, i.e.
   \[
   \text{Cov}[\varepsilon_i, \varepsilon_j] = \begin{cases} 
   0 & \text{if } i \neq j \\
   \text{Var}[\varepsilon_i] & \text{if } i = j
   \end{cases}
   \]
5. The errors are normally distributed

### 3.1.3 Ordinary Least Squares

A common method for estimating the regression coefficients in (2) and (3) is the method of least squares. The idea of the method is to minimize the sum of squares of the errors, i.e. \( S(\beta) = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta) \), through solving the least squares normal equations and ultimately arrive at the following least squares estimator of \( \beta \): [32, p. 70-73]

\[ \hat{\beta} = (X'X)^{-1}X'y \] (4)

As above definition implies, the least squares estimator only exists if the inverse of the matrix \( X'X \) exists, i.e. if the regressor variables are linearly independent. The least squares estimator is normally preferred due to its statistical properties. By the Gauss-Markov theorem, \( \hat{\beta} \) is the linear minimum-variance unbiased (\( E[\hat{\beta}] = \beta \)) estimator, where \( \text{Var}[\hat{\beta}] = \sigma^2(X'X)^{-1} \). [32, p. 79-80]

Using this estimator, the fitted regression model (5) and the residual (6) can now be expressed as below: [32, p.73]

\[ \hat{y} = X\hat{\beta} \] (5)

\[ e = y - \hat{y} \] (6)
3.2 Model Adequacy

Model adequacy checking is necessary in order to determine if any of the initial assumptions made prior to performing the regression are violated. If there is any indication of violation then the model may be unstable and it is likely that the model would differ when regressing over different data samples. There are several different methods for checking if any of the five main assumptions are violated, two being residual rescaling and residual plots. [32, p.131]

3.2.1 Residual Rescaling

To begin with, rescaling of residuals is useful for identifying potential outliers or extreme values. Outliers are observations that are remote in either the regressor or the response space. Furthermore, outliers that lack a subject-specific motivation might point out flaws in the model, for instance regions of the regressor space where the model fits the data poorly. The methods for scaling residuals used in R are standardized residuals and R-student. They are defined as the following, the first one being the standardized residual and the second one being the R-student: [32, p. 129-130]

\[ d_i = \frac{e_i}{\sqrt{\text{MS}_{\text{Res}}}}, \quad i = 1, \ldots, n \quad (7) \]

\[ t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}}, \quad i = 1, \ldots, n \quad (8) \]

Where \( e_i \), \( \text{MS}_{\text{Res}} \) and \( S^2_{(i)} \) is the residual, the estimation of the approximate average variance based on all the observations and the estimation of the approximate average variance based on all the observations except the ith observation respectively. \( \text{MS}_{\text{Res}} \) is defined as: [32, p. 130]

\[ \text{MS}_{\text{Res}} = \frac{\sum_{i=1}^{n} (e_i - \bar{e})^2}{n - p} \quad (9) \]

\( S^2_{(i)} \) is furthermore defined as: [32, p. 135]

\[ S^2_{(i)} = \frac{(n - p)\text{MS}_{\text{Res}} - e_i^2/(1 - h_{ii})}{n - p - 1} \quad (10) \]

If the model is correct, i.e. if the initial assumptions hold, the standardized residuals should have mean zero and a variance that is approximately equal to unity. A standardized residual that is above 3 indicates that the observation is a potential outlier. [32, p. 133]

3.2.2 Residual Plots

Another method for investigating whether the initial assumptions hold is to produce and analyze residual plots. Beyond indicating violations of these assumptions, residual plots also serve as an indicator of measures to amend these violations. Furthermore, there are a large number of different residual plots that can be considered. This thesis will include the normal probability plot, plots of residuals against fitted values, plots of residuals against regressors and the partial regression plots, among others. [32, p. 136]
The normal probability plot can be used to investigate whether the assumption relating to normally distributed errors has been violated. The cumulative normal distribution is plotted as a straight line with R-student residuals on the y-axis and theoretical quantiles on the x-axis. A graph where the residuals approximately follow the straight line is an indicator of normally distributed observations. Further on, plots of residuals against fitted values and regressor values represent the variation of the residuals. The results can be used in order to determine whether the assumption regarding constant variance of the error terms has been violated. Ideally, the residuals should be contained within a horizontal span, indicating that the variance of the residuals is constant. [32, p. 136-139]

Furthermore, partial regression plots are helpful in investigating whether the relationship between the regressor and the response is in fact linear. The partial residuals for the response $y$ and the regressor $x_j$ are plotted against each other. A plot containing a straight line is an indication of a linear relationship between the response $y$ and regressor $x_j$. [32, p. 143-144]

### 3.2.3 Leverage and Influence

In regression analysis, different observations have different amounts of impact or influence on model properties and outcome. Observations that are remote in terms of regressors but not in responses are called leverage points and affect model properties such as $R^2$ but not necessarily parameter estimates. Observations that are remote in terms of both regressor and response are called influence points and affect the model as a whole by shifting it towards the influential observation’s own coordinates. Ensuring that certain observations are not exerting a disproportionate or unjustified amount of influence is important since influential points may for example result in regressors with theoretical importance to be deemed statistically insignificant or regression coefficients with a sign that does not make sense. Consequently, the procedure of detecting and diagnosing such influential points is an important part of model development. [32, p. 211-212]

#### 3.2.3.1 Measure of Leverage

The amount of leverage that the $i$th observation exerts on the $i$th fitted value $\hat{y}_i$ is equal to the standardized distance from the center of the regressor space to the $i$th observation and defined by the diagonal element $h_{ii}$ of the hat matrix $H = X(X'X)^{-1}X'$, which can be expressed as below:

$$h_{ii} = x'_i(X'X)^{-1}x_i$$  \hspace{1cm} (11)

where $x'_i$ is the $i$th row of the matrix $X$. In words, large values of $h_{ii}$ correspond to observations with high leverage and thus potentially high influence. In order to determine whether a value is to be deemed large, the average value of $h_{ii}$ is used as benchmark:

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{\text{Rank}(H)}{n} = \frac{\text{Rank}(X)}{n} = \frac{p}{n}$$  \hspace{1cm} (12)

The $i$th observation is a high leverage point if $h_{ii} > 2\bar{h}$. Since leverage does not necessarily imply influence, the combination of high leverage and large residuals is used as indication of influence. [32, p. 212-213]
3.2.3.2 Measure of Influence

Cook’s Distance

One of the most common measurements of influence is Cook’s distance. It is defined by the squared distance between the estimation of the regression coefficient based on all n observations, \( \hat{\beta} \), and that based on n-1 observations, i.e. after excluding the ith observation, \( \hat{\beta}(i) \):

\[
D_i = \frac{(\hat{\beta}(i) - \hat{\beta})'X'X(\hat{\beta}(i) - \hat{\beta})}{p\text{MS}_\text{Res}} = \frac{(\hat{y}(i) - \hat{y})'X'X(\hat{y}(i) - \hat{y})}{p\text{MS}_\text{Res}}
\]

The ith observation is influential if \( D_i > 1 \). [32, p. 215-216]

DFBETAS

Another measurement of influence is DFBETAS, which is a statistic that measures the change between the estimate \( \hat{\beta}_j \) based on all n observations and the estimate \( \hat{\beta}_j(i) \) based on all observations except for the ith one in standard deviation units:

\[
\text{DFBETAS}_{j,i} = \frac{(\hat{\beta}_j - \hat{\beta}_j(i))}{\sqrt{S(i)^2C_{jj}}} = \frac{r_{j,i}}{\sqrt{r_{j,j}^t} \sqrt{1 - h_{ii}}} \frac{t_i}{t_i}
\]

where \( R = (X'X)^{-1}X' \).

If \( |\text{DFBETAS}_{j,i}| > \frac{2}{\sqrt{n}} \), then the ith observation has significant influence on the jth regression coefficient and should thus be examined more closely. [32, p. 217]

DFFITS

Instead of measuring a certain observation’s influence on a certain regression coefficient, the statistic DFFITS\(_i\) measures the ith observation’s influence on the ith fitted or predicted value. In other words, the statistic is an indicator of both prediction error and leverage and by definition equal to the change in the fitted value if the ith observation is excluded measured in standard deviation units, i.e.:

\[
\text{DFFITS}_{i} = \frac{\hat{y}_i - \hat{y}(i)}{\sqrt{S(i)^2 h_{ii}}} = \sqrt{\frac{h_{ii}}{1 - h_{ii}}} * t_i
\]

If \( |\text{DFFITS}_{i}| > \frac{2\sqrt{p}}{\sqrt{n}} \), then the ith observation has significant influence on the ith predicted or fitted value and should be examined. [32, p. 217-218]

3.2.3.3 Measure of Model Performance

\( D_i \), DFBETAS\(_{j,i}\), and DFFITS\(_i\) are useful in analyzing the effect of individual observations on regression estimates \( \hat{\beta}_j \) and fitted values \( \hat{y}_i \) but not in determining the performance and precision of the model in general. For this purpose, the impact of the ith observation on the estimation precision is defined as follows:

\[
\text{COVRATIO}_i = \frac{|(X'(i)X(i))^{-1}S_{(i)}^2|}{|(X'X)^{-1}\text{MS}_\text{Res}|}
\]

16
If COVRATIO$_i$ > 1, then the $i$th observation helps to better the precision of estimation. If COVRATIO$_i$ < 1, on the other hand, then the $i$th observation worsens the precision. The cut-off values for COVRATIO$_i$ is that it should be either above $1 - \frac{3p}{n}$ or below $1 + \frac{3p}{n}$. [32, p. 219]

3.3 Model Properties

3.3.1 Multicollinearity

Multicollinearity refers to the issue of near-linear dependencies among the regressor variables. If multicollinearity is present in the model, then the least squares estimated regression coefficients can be unstable. The unstability is caused by the inflation of the variance of the least-squares estimators that occurs when multicollinearity is present. [32, p. 288-290]

Mathematically, multicollinearity is defined in terms of the linear dependence among the columns of the $X$ matrix. The columns of the $X$ matrix, denoted $X_j$, are linearly dependent if there exists any $t_1, t_2, ..., t_p$, not all zero, such that: [32, p. 219]

$$\sum_{j=1}^{p} t_j X_j = 0 \quad (17)$$

The problem of multicollinearity is said to be present if equation (17) holds approximately, i.e. if the columns of the $X$ matrix are approximately linearly dependent. [32, p. 219]

Furthermore, multicollinearity can be detected by inspection of the correlation matrix, the variance inflation factor (VIF) and the condition number. Inspection of the correlation matrix refers to examining pairwise correlations between different regressor variables. In practice, this is performed by considering the off-diagonal elements in the correlation matrix. If the off-diagonal elements are close to unity, then the two regressor variables are nearly-linear dependent. [32, p. 292-294]

Moreover, the VIF is defined as: [32, p. 296]

$$VIF_j = (1 - R^2_j)^{-1} \quad (18)$$

where $R^2_j$ is the coefficient of multiple determination when regressing $x_j$ on the remaining regressor variables. If $x_j$ is linearly dependent on a subset of the other regressor variables then $R^2_j$ will be large and consequently the VIF will be large. A VIF that exceeds 10 indicates that the estimation of the associated regression coefficient is inflated due to multicollinearity. [32, p. 296]

The condition number, on the other hand, is defined in terms of the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ of the correlation matrix: [32, p. 298]

$$\kappa = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \quad (19)$$

A condition number below 100 indicates that there is no issue related to multicollinearity present in the model. Condition numbers between 100 and 1000 indicate that the issues are moderate and a condition number that exceeds 1000 indicates that there are severe problems related to multicollinearity. [32, p. 298]
### 3.3.2 Endogeneity

Endogeneity refers to violation of the assumption that the mean of the error term is equal to zero, i.e. \( \mathbb{E}[\varepsilon_i] = 0 \), because the value is dependent on, and thus correlates with, one or more of the regressors. Mathematically, this phenomenon implies that the ordinary least-squares estimation will not result in unbiased estimates. Endogeneity can furthermore have a number of different causes but the most common ones are missing regressor variables, measurement errors, simultaneity and sample selection bias. [30]

### 3.3.3 Heteroscedasticity

Heteroscedasticity refers to violation of the assumption that the variance is constant. The source of this violation is normally the fact that the response follows a non-normal distribution, in which the variance can be expressed as a function of the mean, for instance the Poisson distribution. If not amended, the least squares estimators will remain unbiased but no longer have minimum variance, which implies that the standard deviation will be unnecessarily elevated. Given this potentially rather severe consequence, it is important to detect the phenomenon of heteroscedasticity. The main method for this is to analyze residual plots, as mentioned in the model adequacy section. [32, p. 172-175]

### 3.3.4 Analysis of Variance

Analysis of variance is normally used to derive the test statistic \( F_0 \) used in the hypothesis tests for significance of regression, described in section 3.3.6.1. The total sum of squares, \( SS_T \), is equal to the sum of the sum of squares due to regression, \( SS_R \), and the residual sum of squares, \( SS_{Res} \). [32, p. 84-85]

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square Error</th>
<th>( F ) Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>( SS_R = \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 )</td>
<td>( k )</td>
<td>( MS_R = \frac{SS_R}{k} )</td>
<td>( F_0 = \frac{MS_R}{MS_{Res}} )</td>
</tr>
<tr>
<td>Residual</td>
<td>( SS_{Res} = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2 )</td>
<td>( n-k-1 )</td>
<td>( MS_{Res} = \frac{SS_{Res}}{n-k-1} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( SS_T = \sum_{j=1}^{n}(y_j - \bar{y})^2 )</td>
<td>( n-1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3.3.5 Coefficient of Determination

One application of the information gained from the analysis of variance is to assess the global model adequacy using the coefficient of determination. For this purpose, there are in fact two alternative statistics: \( R^2 \) and \( R^2_{adj} \) adjusted.

\[
R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} \quad (20)
\]

\[
R^2_{adj} = 1 - \frac{SS_{Res}/(n-k-1)}{SS_T/(n-1)} \quad (21)
\]

Analytically, \( R^2 \) is the share of the variance in the response variable that is in fact explained by the regression model. However, \( R^2 \) is always increased by adding more regressor variables without taking into consideration whether the added variable actually improves the model. Hence, there is a risk of overfitting the model, which should always be considered in the model development process. For this reason, \( R^2 \) adjusted is often preferred, since this statistic only increases if the added regressor variable implies an actual reduction of the residual mean square, \( MS_{Res} \). [32, p. 87-88]
3.3.6 Hypothesis Testing

Apart from the five assumptions stated in section 3.1.2, hypothesis testing requires the additional assumption that the errors be independent. This assumption is, however, implied when the errors are both uncorrelated and normally distributed. [32, p. 129] Hypothesis testing has two main purposes: Testing the regression model as a whole, which is commonly referred to as testing for significance of regression, and testing individual, or a subset of, regression coefficients. [32, p. 84-88]

3.3.6.1 F-Statistic

In conformity with analysis of the coefficient of determination, a test for significance of regression is considered a global test of model adequacy. The goal is furthermore to investigate whether there is in fact a linear relationship between the response variable and at least one of the regressors. This test can hence be described using the following hypotheses:

\[ H_0 : \beta_0 = \beta_1 = ... = \beta_k = 0 \]
\[ H_1 : \beta_j \neq 0 \text{ for at least one } j \]

To test the null hypothesis, the F-statistic derived through analysis of variance is used. The null hypothesis is rejected if \( F_0 > F_{x,k,n-k-1} \). [32, p. 84-85] As mentioned in the introduction to this section, the same test can be performed for a subset of regressors, only with a slightly altered F-statistic, which is referred to as a partial F-test. [32, p. 90]

3.3.6.2 t-Statistic

Once confirmed that at least one regressor variable is significant in explaining the behavior of the response, it is natural to investigate which one(s) can be excluded from the model. This is important since the addition of a regressor variable in a model always increases the regression sum of squares but also the variance of the fitted value, \( \hat{y} \), which of course ideally should be minimized. For testing the significance of a specific regressor, the hypotheses are formulated as below:

\[ H_0 : \beta_j = 0 \]
\[ H_1 : \beta_j \neq 0 \]

The null hypothesis is rejected if \( |t_0| > t_{x/2,n-p} \), where the t-statistic is defined as:

\[
t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\sqrt{SE(\hat{\beta}_j)}} \tag{22}
\]

and \( C_{jj} = \text{diag}((X'X)^{-1}) \), \( j = 1, \ldots, n \) and SE is the abbreviation for standard error. In words, the test of significance of an individual regressor is a marginal test of the contribution of the specific regressor, given that the other regressors are present in the model. [32, p. 88]

3.3.7 Confidence Intervals

In conformity with hypothesis testing, confidence intervals require the additional assumption that the errors be independent. In words, a confidence interval with confidence level 100(1 – \( \alpha \)) states that the true value of the estimated parameter will lie within the interval in 100(1 – \( \alpha \)) percent of cases. For an estimated regression coefficient \( \hat{\beta}_j \), a 100(1 – \( \alpha \)) confidence is defined mathematically as: [32, p. 97-100]
\[ \hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \] (23)

### 3.4 Biased Regression

A common consequence of multicollinearity is that the variation of the least squares estimates \( \hat{\beta} \) of the regression coefficients is inflated. This ultimately leads to least squares estimates that are potentially unstable. Essentially, the problem originates from the requirement that the estimate \( \hat{\beta} \) be an unbiased estimate of the actual \( \beta \). The Gauss Markov property states that the least squares estimate has the smallest variance of all the unbiased linear estimates. However, this does not imply that the variance is in fact small. In order to amend the inflated variance, biased regression can be performed. [32, p. 305]

The idea behind biased regression is that by dropping the requirement of an unbiased estimate, one could supposedly find a biased estimate \( \beta^* \) such that the mean squared error, i.e. the sum of the variance and the square of the bias, of \( \beta^* \) is lower than the variance of \( \hat{\beta} \). [32, p. 305] Two procedures to obtain biased estimates, both with the goal of reducing the variance of the coefficient estimate, are Ridge regression and Lasso regression. These methods differ in their penalty term, where Ridge regression has an elliptical constraint and Lasso regression has a polyhedron constraint. As a consequence of the form of the constraints, Ridge regression only shrinks the estimates whilst Lasso regression shrinks as well as performs variable selection as part of the procedure. [17, p. 61-69]

### 3.5 Variable Selection and Model Building

In theory, a common assumption is that all regressor variables are known to be significant. In practice, this is rarely the case. Instead, the number of candidate regressors is normally large but only a select few are shown to be significant and will be represented in the final model. Consequently, the problem of selecting the optimal subset of regressor variables is central but also difficult due to two conflicting objectives – the larger the number of regressors the more information available in order to predict somewhat accurate responses but also the larger the variance of the predicted response. [32, p. 327-328]

#### 3.5.1 Method

In selecting variables, it is desirable to consider and analyze models containing different subsets of regressor variables. There are, however, several possible computational techniques or methods for generating these subsets to choose from. If there are less than approximately 30 regressor variables in the full model, the exhaustive but otherwise superior method of all possible regressions can be used. [32, p. 338]

The all possible regressions method implies fitting all regression models consisting of all possible combinations of 1, 2,..., K candidate regressor variables, where K is the number of regressor variables in the full model. This brings that \( 2^K \) models are to be examined and for each number of candidate regressor variables, the best model is selected, where best signifies smallest \( SS_{Res} \) or highest \( R^2 \). Consequently, the method is computationally demanding for large number numbers of candidate regressor variables and serves as to reduce the number of possible models from \( 2^K \) to \( p = K+1 \). In order to choose the best model out of these \( p \) candidate models, the model measurements introduced in section are 3.5.2 are used. [32, p. 338]
3.5.2 Criteria

When the number of candidate models have been reduced to \( p \), measurements of training error such as \( \text{SS}_{\text{Res}} \) or \( R^2 \) can no longer be used seeing as both favor an increasing number of regressor variables and would consequently recommend the full model. Instead, it is the test error that is desired and as a result, two options remain. First and foremost, it is possible to adjust measurements of training error to take testing into account. Examples of such measurements are the adjusted coefficient of determination, residual mean square error, Mallow’s \( C_p \) and BIC. Secondly, it is possible to directly estimate the test error by excluding subsets of the training observations during the fitting process. This procedure is normally referred to as cross-validation. [23, p. 210] In this analysis, both approaches will be applied, starting with adjusted measurements.

3.5.2.1 Coefficient of Determination (\( R^2_{p} \))

The coefficient of multiple determination is defined as in section 3.3.5, with the only difference being that \( p \) is used to underline that the model has \( p \) coefficients, i.e. \( p \)-1 regressor variables. As mentioned in the model adequacy section, the coefficient \( R^2_p \) increases concavely with the number of regressors, regardless of the value or information added by the regressor. To avoid this issue, the adjusted statistic is normally used, which takes into consideration whether or not the information contained in the added regressor variable is enough to justify its presence in the model: [32, p. 332-333]

\[
R^2_{\text{adj},p} = 1 - \frac{n-1}{n-p}(1 - R^2_p) \tag{24}
\]

3.5.2.2 Residual Mean Squared Error (\( \text{MS}_{\text{Res}}(\!p\!) \))

Unlike the coefficient of determination, the residual mean squared error is a convex function of \( p \) and is defined as below:

\[
\text{MS}_{\text{Res}}(\!p\!) = \frac{\text{SS}_{\text{Res}}(\!p\!)}{n-p} \tag{25}
\]

This means that as \( p \) increases, \( \text{MS}_{\text{Res}}(\!p\!) \) decreases until the point where the reduction in \( \text{SS}_{\text{Res}}(\!p\!) \) is less than the increase in the denominator from losing one degree of freedom. By plotting \( \text{MS}_{\text{Res}}(\!p\!) \) versus \( p \), the value of \( p \) that minimizes \( \text{MS}_{\text{Res}}(\!p\!) \) can easily be observed. Furthermore, it can be proven that the coefficient of multiple determination is equivalent to the residual mean square by the following: [32, p. 333-334]

\[
R^2_{\text{adj},p} = 1 - \frac{n-1}{n-p}(1 - R^2_p) = 1 - \frac{\text{SS}_T}{n-1} \tag{26}
\]

Worth noting is that the residual mean square of the training set is normally an underestimate of that of the testing set, since the regression coefficient estimates are chosen as to minimize the training residual mean square error. For this reason, a number of techniques to adjust the training error have been brought forward. These are presented in the following subsections.[23, p. 210]
3.5.2.3 Mallow’s ($C_p$)

The statistic Mallow’s $C_p$ is an estimate of the test mean squared error and defined as: [23, p. 211]

$$C_p = \frac{SS_{Res}(p)}{\hat{\sigma}^2} - n + 2p$$  \hfill (27)

Regression models without any bias have $C_p = p$. Consequently, the statistic is both a measurement of prediction error and bias since a low value of the statistic implies low prediction error whilst a value close to $p$ implies low bias. In many cases, a low prediction error is more important than the level of bias in the model. [32, p. 334-336]

3.5.2.4 Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC), on the other hand, is a log-likelihood measure that places a penalty on adding regressors. There are however a number of versions of the BIC. The version used in R is the Schwartz version, defined as: [32, p. 334-336]

$$BIC_{Sch} = n \times \ln(\frac{SS_{Res}}{n}) + p \times \ln(n)$$  \hfill (28)

3.5.2.5 Cross-Validation

Cross-validation entails dividing the original sample of observations into $k$ different samples of equal size. Each of the $k$ samples are in turn treated as a validation sample and the model is fitted on the observations in the remaining $k-1$ samples. For each such fitting process, the mean squared error is calculated on the observations in the validation sample, resulting in $k$ different values of the mean squared error. The cross-validation estimate of the test error is obtained by taking the mean of these $k$ values. [23, p. 176-183]

3.5.3 Strategy for Model Building

In order to solve the problem of variable selection and model building, an iterative approach and strategy is required. It can be summarized as: [32, p. 327-328]

1. Fit the full model
2. Perform residual analysis
3. Determine if transformation is necessary
4. Perform variable selection
5. Select a set of best models
6. Check model adequacy for this set
7. Make recommendations
4 Methodology

4.1 Literature Study

In order to understand the topic of macroeconomic variables and their impact on stock markets, macroeconomic theory has been studied and a literature review of existing research within the field has been performed. Both procedures have been valuable in choosing not only regressor variables but also what measures of the variables to use, since most macroeconomic variables can be measured in a number of different ways. The main literature used for macroeconomic theory is Macroeconomics (Dornbusch et al, 2018). In reviewing existing literature and results specifically within the field macroeconomic variables’ impact on stock markets, the main sources of research have been the Library of the Royal Institute of Technology (KTH) and Google Scholar. Key words used are “macroeconomics”, “macroeconomic variables”, “stock return” and “regression analysis”. The articles examined more closely were selected based on title, abstract, journal and time of publication.

The main literature used for the mathematical theory on regression analysis is Introduction to Linear Regression (Montgomery et al, 2012) and An Introduction to Statistical Learning (James et al, 2013).

4.2 Data Collection

The data sample used for the multiple regression analysis of this thesis consists of 213 monthly observations during the period 2001:M1-2018:M09. Furthermore, it consists of public data only. Stock return data has been collected from the Nasdaq Nordic website, whilst all macroeconomic data has been collected from public authorities such as the Central Bureau of Statistics (Statistiska Centralbyrån) and the National Institute of Economic Research (Konjunkturinstitutet) or secondary sources such as the Confederation of Swedish Enterprises’ (Svenskt Näringslivs) own website Ekonomifakta, which presents statistics originally made available by the Central Bureau of Statistics. [9]

4.3 Data Processing

Initially, the data was processed and compiled in order to arrive at a homogeneous scale. The stock price data from the Nasdaq Nordic website consists of daily closing prices whilst GDP is published quarterly and most other macroeconomic variables monthly. Since the majority of variables of interest are published monthly, this analysis uses data points of a monthly scale.

As a consequence, the variables with data published with another frequency, stock return and GDP in this case, must be processed. To begin with, the stock return data was processed by taking the mean of all daily closing prices of each month, which excludes all fluctuations that take place during the day. The processing is enabled by the assumption that the removed fluctuations will not affect the analysis greatly since performed over a relatively long time period. When it comes to GDP published quarterly, it was assumed that GDP is constant during the whole quarter in order to receive monthly data points.

When all variables consisted of monthly data points, all data sets were processed in excel in order to create one single data sample to be read into R, where the actual analysis was be performed.
4.4 Analysis

The chosen method is multiple linear regression analysis, which offers an intuitive and practicable approach to modelling linear relationships between variables. As a consequence, the method was chosen under the assumption that the relationships between the variables are in fact linear - which may not necessarily be true. For this reason, the linearity assumption will be tested in the analysis section and re-evaluated if the results indicate that it should be necessary.

Furthermore, it should be mentioned that multiple linear regression implies correlation but not necessarily causation. This aspect is deemed relevant to consider since the macroeconomic relationships are complex and the use of a different method, for example vector autoregression, may well have lead to different results. This fact is acknowledged and will be discussed in later sections. Although other methods may imply certain advantages, multiple linear regression is used given the scope of the thesis, the feasibility of the method and the fact that multiple linear regression has been used in previous studies within the field.

5 Results

5.1 Full Model

5.1.1 Full Model Properties

5.1.1.1 Bank Index

Table 1 presents the summary statistics of the bank index, which includes the estimates of the regression coefficients for the full model, their corresponding standard error as well as the value of the t-statistic for each coefficient. The $R^2$ as well as the $R^2_{adj}$ are high, indicating that the model encompasses a large proportion of the variance in the response. Further on, the F-statistic shows that the full model is significant, which implies that it contains at least one regression coefficient which is non-zero. By the t-tests it can be concluded that all of the individual variables except for consumer expectations and repo rate are found significant. This suggests that these two variables may be redundant and consequently that variable selection should be considered.

|               | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 9576.1228| 1002.4372  | 9.55    | 0.0000   |
| Unemployment  | -34.4328 | 17.0330    | -2.02   | 0.0445   |
| Money Supply  | 0.0012   | 0.0001     | 10.49   | 0.0000   |
| Repo rate     | -28.6538 | 19.3840    | -1.48   | 0.1409   |
| Inflation     | -32.4431 | 3.2993     | -9.83   | 0.0000   |
| GDP           | 0.0010   | 0.0002     | 4.33    | 0.0000   |
| Exchange dollar| -48.4489| 18.5147    | -2.62   | 0.0095   |
| Exchange euro | -148.2691| 23.9447    | -6.19   | 0.0000   |
| Consumer expectations | 2.3307  | 1.4160     | 1.65    | 0.1013   |
| Oil price     | 5.4872   | 1.0091     | 5.44    | 0.0000   |

F-statistic: 109  p-value: < 2.2e-16
Multiple R-squared: 0.8285
Adjusted R-squared: 0.8209

Table 1: Summary Statistics - Bank Index
5.1.2 Full Model Properties

5.1.2.1 Industry Index

The corresponding summary statistics for the industry index are presented in Table 2 below.

|                  | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------|----------|------------|---------|---------|
| (Intercept)      | 1659.1512| 478.6735   | 3.47    | 0.0006  |
| Unemployment     | -7.1336  | 8.1334     | -0.88   | 0.3815  |
| Money Supply     | 0.0010   | 0.0001     | 18.61   | 0.0000  |
| Repo rate        | 45.9368  | 9.2560     | 4.96    | 0.0000  |
| Inflation        | -10.5273 | 1.5754     | -6.68   | 0.0000  |
| GDP              | 0.0003   | 0.0001     | 3.08    | 0.0024  |
| Exchange dollar  | -11.3183 | 8.8409     | -1.28   | 0.2019  |
| Exchange euro    | -26.2753 | 11.4338    | -2.30   | 0.0226  |
| Consumer expectations | 6.8247 | 0.6761     | 10.09   | 0.0000  |
| Oil price        | 2.3929   | 0.4818     | 4.97    | 0.0000  |

F-statistic: 682.5  p-value: < 2.2e-16
Multiple R-squared: 0.968
Adjusted R-squared: 0.9666

Table 2: Summary Statistics - Industry Index

The high $R^2$ and $R^2_{adj}$ values indicate that the model accounts for almost all variance in the response. The F-test furthermore shows that the global model is significant and the t-tests indicate that all variables, except possibly for unemployment and the SEK/USD exchange rate are significant.

5.1.3 Residual Analysis

To begin with, the partial regression plots, represented below as Figure 1 and Figure 2, are examined, where each plot illustrates the relationship between the response and each of the regressors. The plots are helpful in determining whether the assumption of a linear relationship between the response and the regressors may have been violated.
For the bank index the partial regression plots found in Figure 1 above illustrate a straight line for all variables and therefore the relationship between the response and each regressor is assumed to be linear if existent. However, the two plots corresponding to regressors consumer expectation and repo rate respectively present a rather flat line. This indicates that the response variable may well be independent of these two variables, which is coherent with the results from the t-tests presented earlier.
Moreover, similar conclusions can be drawn for the *industry index*, found in Figure 2 above. All partial regression plots illustrate a linear relationship between response and each of the regressors, except possibly for the plots corresponding to *unemployment* and *SEK/USD exchange rate*. Both plots show a rather flat line, indicating that they might be redundant in the model. This conclusion is furthermore supported by the t-test presented in the model properties section.

Further on, the residual versus regressor value plots are examined and presented for both models in Figure 3 on the next page. In general, it holds that if the residuals can be contained within a horizontal band, then no severe violations against the assumptions are indicated. A clear outward-opening funnel or double-bow pattern, on the other hand, would indicate that the variance is non-constant and increases as x increases. Furthermore, a curved plot would indicate that the assumption of linearity is violated. Since Figure 3 does not illustrate any funnel or double-bow patterns but rather that most of the residuals can be contained within a horizontal band, homoscedasticity is assumed. When it comes to the linearity assumption, the results are less clear for the *bank index*, since the left plot is not completely linear. However, despite this fact, the graph is still rather flat and for this reason it assumed that the linearity assumption has not been violated.

![Residuals vs Fitted Values](image)

(a) Bank Index  
(b) Industry Index

Figure 3: Standardized Residuals vs Fitted Values

The residual vs regressor plots generally indicate similar assumption violations as the residuals vs fitted values and are analyzed with regard to the same criteria. The graphs are presented in Figure 4 and 5 below, where each plot corresponds to a specific regressor. Since none of the plots illustrate a curved plot, linearity is assumed. Furthermore, more or less all residuals are contained within a horizontal band, indicating that homoscedasticity can be assumed.
Moreover, the assumption that the errors are normally distributed is considered. This is performed by examining the normality plot for both models, presented in Figure 6 below.
The plots in Figure 6 clearly illustrate that the residuals more or less follow a straight line, indicating that the normality assumption holds. It is, however, apparent from the plots that the curves are slightly skewed, that is to say that they illustrate an upward curve in the upper-right corner. This indicates that the tails of the distributions are slightly lighter than the tails corresponding to a normal distribution. However, according to Montgomery et al (2012), slight deviations from the straight line, such as these ones, do not invalidate the normality assumption and therefore normality is assumed.

5.1.4 Analysis of Outliers

In this section, both leverage and influence of data points will be considered and analyzed. To begin with, leverage specifies whether an observation is considered an outlier in the regressor space whilst influence specifies whether an observation is considered an outlier in both regressor and response space. More specifically, the plot of residuals vs leverage is commonly used as an indicator of leverage. As presented in the graphs to the left of Figures 7 and 8, some observations exceed the cut-off value of leverage, marked with red. However, since this does not necessarily imply that the observations are influence points, the measure of Cook’s distance is used. These results are presented in the two middle graphs of Figures 7 and 8. As illustrated, none of the observations exceed the cut-off value for Cook’s distance, which was found to be approximately 0.9373. Last but not least, the graphs to the right in Figures 7 and 8 furthermore illustrate that there are observations with high levels of leverage but low levels of influence, which implies that there are observations that are remote in regressor space but that the corresponding response value is close to the fitted value. Consequently, the outliers present in both models are generally leverage points rather than influence points and thus not exerting undue influence on model properties.
Secondly, the COVRATIO, DFBETAS and DFFITS diagnostics are examined and presented below in Figures 9, 10 and 11. As illustrated by Figure 9, the majority of the observations for both models lie within the boundaries of the cut-off values for COVRATIO, indicated by the red lines. This result suggests that the addition of observations does not increase the variance of the estimated regressor coefficients. As presented by Figure 10, most observations furthermore lie within the cut-off values for DFFITS, indicating that the fitted values are not greatly influenced by particular observations. Lastly, Figure 11 presents a histogram of DFBETAS for both models. The graphs illustrate that most observations only marginally affect regressor coefficients. From the combination of these results, it can be concluded that exclusion of specific observations would not noticeably affect regressor coefficients and consequently that influence is not an issue in either of the models.

Although some observations exceed the cut-off values for the DFFITS, DFBETAS and the COVRATIO diagnostics, and consequently might have an overly influential effect on the model, they have not been removed from the model. The decision to include all the observations is based on the fact that none of the observations are invalid in the sense that they have been measured incorrectly or do not belong to the data-population of interest (e.g. belongs to a year that is not within the range 2001-2018).
Figure 9: Analysis of Influence - COVRATIO

Figure 10: Analysis of Influence - DFFITS

Figure 11: Analysis of Influence - DFBETAS
5.1.5 Multicollinearity

Initially, it is worth clarifying that this analysis of multicollinearity does not distinguish the two models from each other, seeing as the analysis is independent of response and regressors are common. Furthermore, the analysis is based on three fundamental diagnostics: The correlation matrix, the variance inflation factor (VIF) and the condition number. To begin with, the VIF values are presented in Table 3 for all regressor variables. As presented, the variables money supply and inflation have VIF values that exceed the cut-off value of 10, which indicates that the variance of these two coefficients is significantly inflated due to multicollinearity.

<table>
<thead>
<tr>
<th></th>
<th>Unemployment</th>
<th>Money Supply</th>
<th>Repo rate</th>
<th>Inflation</th>
<th>GDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>2.84</td>
<td>36.39</td>
<td>8.41</td>
<td>30.66</td>
<td>5.68</td>
</tr>
<tr>
<td></td>
<td>SEK/USD</td>
<td>SEK/EUR</td>
<td>Consumer Expectations</td>
<td>Oil Price</td>
<td></td>
</tr>
<tr>
<td>VIF</td>
<td>4.02</td>
<td>1.27</td>
<td>1.44</td>
<td>6.25</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Variance Inflation Factor

Secondly, a representation of the correlation matrix is presented in Figure 12. The matrix illustrates pair-wise correlations between different regressors and a red colour indicates a negative correlation whilst a blue colour indicates a positive correlation. Consequently, it can be concluded that GDP, money supply and inflation all have high positive pair-wise correlations, which is not completely unexpected considering presented economic theory. Similarly, the regressor variable pairs repo rate and GDP, money supply and inflation, respectively, are associated with a negative correlation.

![Figure 12: Correlation Matrix](image-url)
Although the correlation matrix only represents pair-wise relationships, and consequently is not a sufficient diagnostic analyzing multicollinearity involving more than two regressors, it is worth acknowledging that the VIF and the correlation matrix indicate that the same variables suffer from multicollinearity. Lastly, the condition number is found to be $\frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} = 249.4194$, which indicates that the model as a whole has a moderate but not severe amount of multicollinearity.

### 5.1.6 Lasso regression

One approach to combat the issues related to multicollinearity is to perform biased regression. To this end, there are several different methods available and one of them is Lasso regression, which is used in this analysis. The results for each regression are presented in Figure 13 and the first plot illustrates the mean squared error for different values of the biasing parameter ($\lambda$). The first vertical line represents the $\lambda$ such that the lowest mean squared error is obtained and the second vertical line represents the $\lambda_{1\text{se}}$, i.e., the $\lambda$ one standard deviation from the $\lambda$ corresponding to the lowest mean squared error. The second plot illustrates the estimated regression coefficients for different values of $\lambda$. The number of regressors for different values of $\lambda$ is illustrated by the numbers above the plots.

![Lasso Regression](https://example.com/lasso_regression.png)

Figure 13: Lasso Regression

Based on the $\lambda_{1\text{se}}$, it can be concluded that the Lasso regression for the bank index recommends a model with all the regressors whilst the one for the industry index recommends a model with 8 regressors, where the SEK/USD exchange rate has been removed from the model. Furthermore, whether or not the Lasso regression is actually superior to the ordinary least squares regression can be determined by comparing the mean squared error. The mean square error of the Lasso regression will correspond to the $\lambda_{1\text{se}}$. The results are presented in Table 4 below.
Although the differences are very small, it can be concluded from the table that the full model actually is superior to that suggested by Lasso regression when it comes to estimating the regression coefficients. This fact holds for both models. Therefore, the variable selection procedure will be based on the full models rather than those obtained from the Lasso regression. Furthermore, it is general practice to choose the $\lambda$ so that the coefficients converge towards a stabilizing value, which for both indexes occur at a $\lambda$ corresponding to an MSE that exceeds the MSE of the full models, seen in the second plot in Figure 13, further motivating the performance of variable selection based on the full models. Note also that although the mean squared errors are very large, this is not alarming since the index points, i.e. the responses of the models, generally take on values between 1000-2000.

### 5.2 Variable Selection

Given the rather limited number of regressor variables in the full model, the method of all-possible regression has been implemented, resulting in nine candidate models. These nine candidate models are thereafter compared with regard to the statistics Mallow’s $C_p$, BIC, cross-validation and adjusted $R^2$. The results are presented in Figures 14 and 15 below.

![Figure 14](image)

**Figure 14: All Possible Regression - Bank Index**

From the presented graphs it can be concluded that the two criteria Mallow’s $C_p$ and adjusted $R^2$ as well as cross-validation suggests the full model for the bank index. The BIC criterion, on the other hand, recommends a model where the three regressors unemployment, repo rate and consumer expectations have been removed.
For the industry index, it can be concluded that the criteria Mallow’s $C_p$ and BIC both suggest a model where unemployment and SEK/USD exchange rate have been removed. The adjusted $R^2$ criterion, on the other hand, recommends a model where only unemployment has been excluded whilst the results from cross-validation suggest a model with only five regressors, more specifically a model that includes $GDP$, repo rate, money supply, spot price of oil and consumer expectations.

The individual recommendations of the four different criteria for each of the two models are summarized in the table below. Consequently, there are five candidate models.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Bank Model: Number of Variables</th>
<th>Bank Model: Included Variables</th>
<th>Industry Model: Number of Variables</th>
<th>Industry Model: Included Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted $R^2$</td>
<td>9</td>
<td>GDP, Money Supply, Repo Rate, Inflation, Unemployment Rate, Consumer Expectation, Spot price of oil, SEK/USD Exchange Rate, SEK/EUR Exchange Rate</td>
<td>8</td>
<td>GDP, Money Supply, Repo Rate, Inflation, Consumer Expectation, Spot price of oil, SEK/USD Exchange Rate, SEK/EUR Exchange Rate</td>
</tr>
<tr>
<td>BIC</td>
<td>6</td>
<td>GDP, Money Supply, Inflation, Spot price of oil, SEK/USD Exchange Rate, SEK/EUR Exchange Rate</td>
<td>7</td>
<td>GDP, Money Supply, Repo Rate, Inflation, Consumer Expectation, Spot price of oil, SEK/EUR Exchange Rate</td>
</tr>
<tr>
<td>Mallow’s $C_p$</td>
<td>9</td>
<td>GDP, Money Supply, Repo Rate, Inflation, Unemployment Rate, Consumer Expectation, Spot price of oil, SEK/USD Exchange Rate, SEK/EUR Exchange Rate</td>
<td>7</td>
<td>GDP, Money Supply, Repo Rate, Inflation, Consumer Expectation, Spot price of oil, SEK/EUR Exchange Rate</td>
</tr>
<tr>
<td>Cross-validation</td>
<td>9</td>
<td>GDP, Money Supply, Repo Rate, Inflation, Unemployment Rate, Consumer Expectation, Spot price of oil, SEK/USD Exchange Rate, SEK/EUR Exchange Rate</td>
<td>5</td>
<td>GDP, Money Supply, Repo rate, Consumer expectations, Spot price of oil</td>
</tr>
</tbody>
</table>

### 5.3 Final Model

In choosing two final models – one for bank index and one for industry index – the three different variable selection criteria for all final candidate models presented above have been considered. In addition to the variable selection criteria, the results from the cross-validation, Lasso regression, t-statistics and the partial regression plots have also served as decision basis in selecting the final model.

For the model representing the bank index, the criteria $R^2_{adj}$, Mallow’s $C_p$ and cross-validation all suggest that the final model should include all of the regressors. However, the
BIC criterion suggests a model that excludes the regressors repo rate, unemployment and consumer expectations. Subsequently, the t-statistics indicate that the regressors repo rate and consumer expectations can be removed from the model, and that unemployment only has a significance level of 95%. The partial regression plots for the regressors repo rate, consumer expectation and unemployment are all rather flat, indicating that the regressors should not be included in the model. Taking all these indicators into consideration, the final model for the bank index ix consequently chosen to include the regressors SEK/USD exchange rate, SEK/EUR exchange rate, oil price, GDP, money supply and inflation. Although multiple criteria suggested the full model, the difference in criteria values between the final model and the full model are marginal and thus the final model is deemed as the superior model.

Secondly, for the model representing the industry index, the criteria t-statistics, Mallow’s Cp and BIC all suggest that the regressors unemployment and SEK/USD exchange rate should be removed from the model. The partial regression plots for the regressors unemployment, SEK/USD exchange rate and consumer expectations are furthermore all rather flat, indicating that there is no linear relationship between those regressors and the response. Moreover, the Lasso regression removes the regressor SEK/USD exchange rate and the $R^2_{adj}$ criteria suggests a model where unemployment has been excluded. Lastly, cross-validation recommends a final model that excludes the regressors inflation, unemployment, SEK/EUR exchange rate and SEK/USD exchange rate. Although the results are rather inconsistent, the two regressors unemployment and SEK/USD exchange rate are excluded by more or less all of the methods or criteria. Therefore, the final model representing the industry index includes the regressors repo rate, SEK/EUR exchange rate, consumer expectations, oil price, GDP, money supply and inflation.

Furthermore, in order to ensure that the selected final models are reliable and not violating any model assumptions, a brief and final analysis is performed. More specifically, the summary statistics, the normality plot and a number of multicollinearity diagnostics are considered and examined. To begin with, the summary statistics are analyzed. Tables 5 and 6 below present t-statistics, F-statistics, $R^2_{adj}$ and $R^2$ for both models.

|          | Estimate | Std. Error | t value | $Pr(>|t|)$ |
|----------|----------|------------|---------|------------|
| (Intercept) | 1424.5900 | 445.4073  | 3.20    | 0.0016 **  |
| Money supply | 0.0010  | 0.0000  | 23.77   | 0.0000 *** |
| Repo rate  | 49.4748  | 6.1389   | 8.06    | 0.0000 *** |
| Inflation  | -10.2911 | 1.5038   | -6.84   | 0.0000 *** |
| GDP        | 0.0004   | 0.0001   | 3.43    | 0.0007 *** |
| EUR exchange | -28.0124 | 11.3129  | -2.48   | 0.0141 *   |
| Expectations | 6.9199  | 0.6724   | 10.29   | 0.0000 *** |
| Oil price  | 2.6100   | 0.4265   | 6.12    | 0.0000 *** |

Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1 ' '

F-statistic: 874.3 p-value: < 2.2e-16
Multiple R-squared: 0.9676
Adjusted R-squared: 0.9665

Table 5: Summary statistics - Final Industry Index Model
### Table 6: Summary statistics - Final Bank Index Model

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 10515.0135 | 810.8508 | 12.97 | 0.0000 *** |
| Money supply | 0.0013 | 0.0001 | 15.25 | 0.0000 *** |
| Inflation | -36.1148 | 2.8277 | -12.77 | 0.0000 *** |
| GDP | 0.0009 | 0.0002 | 4.14 | 0.0001 *** |
| USD exchange | -57.6785 | 17.5915 | -3.28 | 0.0012 ** |
| EUR exchange | -152.7980 | 23.3080 | -6.56 | 0.0000 *** |
| Oil price | 5.5990 | 0.8850 | 6.33 | 0.0000 *** |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 2e-16

F-statistic: 159.4  p-value: < 2.2e-16

Multiple R-squared: 0.8228
Adjusted R-squared: 0.8176

The F-statistic for both models indicate that there is in fact a linear relationship between the response and at least one of the regressors. Further on, the t-statistics for both models illustrate that all individual regressors included in the finals models are significant, although with varying significance levels, which are indicated by the number of stars to the right in the table. The $R^2$ and $R^2_{adj}$ for the *industry index* are 0.9676 and 0.9665 respectively, which is slightly lower than the $R^2$ and $R^2_{adj}$ for the full model (0.9666 and 0.9680 respectively). Equivalently, $R^2$ and $R^2_{adj}$ for the final model for the *bank index* are lower than those for full the model but this difference can be considered marginal.

Further on, the multicollinearity of the models is analyzed. The two following tables, Table 7 and 8, present the VIF values for each final model.

<table>
<thead>
<tr>
<th>Money Supply</th>
<th>Repo rate</th>
<th>Inflation</th>
<th>GDP</th>
<th>Euro</th>
<th>Expectations</th>
<th>Oil price</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>22.002</td>
<td>3.686</td>
<td>27.846</td>
<td>5.140</td>
<td>1.237</td>
<td>1.422</td>
</tr>
</tbody>
</table>

Table 7: Variance Inflation Factor - Final Industry Index Model

<table>
<thead>
<tr>
<th>Money Supply</th>
<th>Inflation</th>
<th>GDP</th>
<th>Exchange dollar</th>
<th>Exchang euro</th>
<th>Oil price</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>22.302</td>
<td>22.115</td>
<td>5.388</td>
<td>3.560</td>
<td>1.180</td>
</tr>
</tbody>
</table>

Table 8: Variance Inflation Factor - Final Bank Index Model

Given these results, it is clear that the two regressors *money supply* and *inflation* are both highly affected by multicollinearity. However, it is apparent that the VIF values of all regressors are in fact lower in the final models than in the full models. This is especially true for the regressors *money supply* and *inflation*, given the fact that they have VIF values of 36.39 and 30.66 respectively. Therefore, it is concluded that the issue of multicollinearity is not as severe for the final models as for the full models.

Lastly, it is ensured that the normality assumption is not violated. This is concluded by examining the normality plot, presented in Figure 16 below. Although there is a slight skew, it can be concluded that the normality assumptions still hold for both models.
From an economical standpoint the results imply that the macroeconomic variables money supply, repo rate, inflation, GDP, euro exchange, consumer expectations and oil price are all important in explaining the stock return in the industrial goods and services industry. Likewise, it can be established that the macroeconomic variables money supply, inflation, GDP, dollar exchange, euro exchange and oil price are all important in explaining the stock return in the banking industry. Evidently it can be concluded that the macroeconomic variables affecting the respective indexes differ. Furthermore, the variable coefficients represent the change in the index value for each unit change in each variable respectively. Note however that the results only recognizes relationships between the independent and dependent variables, i.e. it is not possible to determine the causality between the variables.

6 Discussion

6.1 Variables

As presented in the previous section, the majority – but not all – of the original regressor variables are still relevant and present in the two final models. In this section, the presence of each regressor variable and the nature of its relation to stock return will be discussed. Furthermore, the results of this thesis will be compared to both previous research examined in the literature review as well as economic theory.

6.1.1 Gross Domestic Product

According to presented results, GDP is positively correlated with stock return within both banking and industrial goods and services. This result is deemed reasonable since it aligns with both economic theory and previous research within the field. Economic theory stipulates that GDP is a basic measurement of output and furthermore that GDP growth is equivalent to economic growth, which in turn is mirrored by the stock market. Moreover, all previous research has presented a positive relation between stock return and GDP.
6.1.2 Money Supply

In conformity with GDP, money supply is positively related to stock return within both industries, which is supported by economic theory. As presented in the economic theory section, an increase in money supply is used to stimulate the economy and should logically therefore affect the stock market positively. Furthermore, money supply is positively related to output (GDP) by the LM relation, which confirms that the sign of the corresponding coefficient is reasonable. Unfortunately, the relationship between money supply and stock return is not as well-explored as that of many other macroeconomic variables. The reviewed research within this field is moreover contradictory and is consequently not used in evaluating the reasonableness of the nature of the relation.

6.1.3 Repo Rate

Beyond the positive relationship between money supply and output, the LM relation stipulates a negative relation between money supply and interest rate. Furthermore, the repo rate is one of the most important tools of monetary policy and a decrease in the repo rate is performed with the purpose of stimulating the economy and vice versa. In practice, this is performed through increasing the money supply. In other words, the relation between repo rate and stock return should logically be negative. This conclusion is moreover supported by all previous research reviewed.

Contrary to economic theory and results from previous research, a positive relation between repo rate and stock return is presented for industrial goods and services, whilst the variable is not deemed significant for the banking industry. Within the field of linear regression analysis, occasionally, the sign of a regression coefficient contradicts theory or intuition. This phenomenon has four main reasons – other significant regressors are excluded from the model, considerable multicollinearity is present, computational errors have been made or the range of the values of the regressor is too narrow. [32, p. 119-121]

To begin with, it is very reasonable to assume that variables with significant impact on the response are not included in the model, since exclusively macroeconomic variables and no company or industry specific information are represented amongst the regressors. Given that the VIF value of the regressor repo rate is relatively low, multicollinearity is however considered unimportant in explaining the non-intuitive sign of the corresponding regression coefficient. The presence of computational errors is moreover considered irrelevant since the data is collected as presented by the Swedish Riksbank and thereafter not manipulated further. However, that the range of the regressor may be too small is considered highly relevant for the analysis. In the data sample, the values of the repo rate range from -0.0050 to 0.0475, resulting in high variance since this value is dependent of the range of the regressor. Consequently, the narrow range in combination with exclusion of significant variables is considered plausible in explaining the non-intuitive or even illogical sign of the regression coefficient. Furthermore, these reasons may also explain that the variable is not significant for banking, since in practice this implies that the coefficient estimate is set equal to zero and can be caused by an inflated variance.

6.1.4 Inflation

According to presented results, inflation is negatively correlated with stock return within both banking and industrial goods and services. This result is considered reasonable, given
that it aligns with previous research within the field. In economic theory, inflation is defined as the continuous increase in price level and acknowledged as an important target of monetary policy. A decrease in repo rate is used to generate an increase in inflation and vice versa and consequently it may be considered reasonable that inflation be positively related to stock return. On the other hand, economic theory also stipulates that high inflation is volatile and that monetary policy normally strives towards maintaining a low and stable inflation for the welfare of the economy. For this reason, it is considered that both previous research and economic theory supports the conclusion that the negative relation is reasonable.

6.1.5 Consumer Expectations

Consumer expectations are positively related to stock return within industrial goods and services but not significant in explaining stock return within the banking industry. The first result is supported by previous research, even though the amount of research on the subject of consumer expectations and stock return is limited. Furthermore, consumer expectations logically mirror the welfare of the economy, explaining the positive correlation. In addition to mirroring the welfare, consumers are also one of the main determinants of economic welfare [7, p. 26-39], rendering consumer expectations one of the more complex variables subject to this analysis since consumer expectations greatly affect consumption. On this note, it may be considered reasonable to assume that industrial goods and services should be more affected by consumers than the banking industry since, in the end, producing companies are indirectly dependent on consumers’ will to consume their goods. On the other hand, banks are dependent on consumers will to take on loans and mortgages and should consequently also be affected by consumer expectations. In other words, it is possible to argue both for and against the obtained result and therefore no firm conclusion can be drawn regarding its reasonableness.

6.1.6 Spot Price of Oil

As explained in the economic theory section, not only industrial corporations but the economy as a whole is very much dependent on non-renewable fossil fuels such as crude oil. As a consequence, it may be considered logical to assume that increased oil price would decrease stock return, since an increased cost of production results in decreased profits. This theory is furthermore supported by previous research, which stipulates a negative correlation between oil price and stock return.

Contrary to economic theory and previous research, the analysis of this thesis results in a positive correlation between oil price and stock return. In conformity with the case of repo rate, this correlation may be considered non-intuitive or even illogical. For the variable oil price, neither multicollinearity nor the range of the regressor values is considered plausible to explain the results. Furthermore, computational errors are considered unimportant since the data is collected as presented by the U.S. Energy Information Administration and thereafter not manipulated further. By elimination of explanations deemed irrelevant, only the exclusion of important regressors remains. As stated for the repo rate, it is considered reasonable to assume that variables with significant impact on the response are not included in the model, since exclusively macroeconomic variables and no company or industry specific information are represented amongst the regressors.
6.1.7 Exchange Rates

According to the results presented, stock return within banking is negatively correlated with both exchange rates included in the full model whilst stock return within industrial goods and services is negatively related to the SEK/EUR exchange rate but not significantly related to the SEK/USD exchange rate. These results contradict the previous research reviewed, where positive correlations were presented. Unlike many macroeconomic variables, such as inflation and money supply, there is however no obvious or indisputable relationship between stock return and exchange rates within the field of economic theory. Instead, the relation is dependent of the trade deficit of the country using the currency in question. The previous research reviewed mainly analyzed the United States, which is a country with a large and positive trade deficit, [29] which generally favors a strong currency. Sweden, on the other hand, has a relatively large but negative trade deficit, which means that exports exceeds imports. [28] Such countries generally benefit from a low currency, which could explain the negative correlation concluded in this thesis. The results are consequently considered reasonable and reliable.

6.2 Model

Thorough analysis of final models have been performed in order to ensure that no fundamental model assumptions are violated. However, this analysis does not ensure that the models are completely spared from flaws and therefore perfectly reliable. For example, it is worth noting that the R\(^2\) and R\(^2\)\(_{\text{adj}}\) of the industry index are very high. Although R\(^2\) is a measurement of model fit, it is not necessarily desirable to achieve the highest possible value of R\(^2\). On the contrary, an R\(^2\) that is considered abnormally or unnaturally high can serve as an indicator that the model is incomplete in some sense. [32, p. 36] Moreover, determining whether or not the R\(^2\) value is reasonable is a process that requires subject knowledge. Considering the fact that the model does not include relevant factors such as industry or company specific information, the high values of R\(^2\) and R\(^2\)\(_{\text{adj}}\) of the industry index model definitely raise some questions regarding the validity of the model.

On the other hand, there are many possible explanations to why the value of R\(^2\) is unnaturally high. For instance, the phenomenon can be explained by the fact that the model is overfitted or that regressor and response values both exhibit trends over time. [32, p. 36] Furthermore, considering the number of observations and the number of regressors, overfitting the model should not be an issue in this case. This is clearly supported by the fact that R\(^2\)\(_{\text{adj}}\), in coherency with the R\(^2\), is very high. The latter explanation regarding trends over time could, however, possibly account for the high R\(^2\) value of the industry index model. This phenomenon can potentially be combated by performing a time series analysis, which will be discussed later on.

Another limitation of the model is the fact that linear regression cannot establish causality without further confirmation. [32, p. 5] This is of special importance in this thesis since the underlying economic theory suggests that the regressor variables are highly correlated, which for instance can be illustrated by the fact that an increase in money supply is followed by a decrease in repo rate and an increase in inflation. [7, p. 197] These correlations are to some extent illustrated by diagnostics of multicollinearity. However, the problem of multicollinearity can be combated by for example variable selection but the issue of determining appropriate causalities between the variables may in some cases still remain. For instance,
if the two variables *repo rate* and *money supply* exhibit high pairwise correlation and as a consequence, one of the variables is removed through variable selection, then ultimately multicollinearity is decreased. However, no decisive or firm conclusions can be drawn regarding which of these two variables is in fact the one with most influence on the response. Furthermore, since the model does not establish causality it is not possible to determine the cause-and-effect relationships between the independent and dependent variables. In other words, it is very difficult to determine the individual relationships between the regressors and the responses.

Furthermore, the number of observations is rather limited, which could possibly have an effect on the validity of the model. Worth noting, however, is that increasing the number observations is not as straight-forward as for many other applications, seeing as the variables are measured over time. This, in turn, implies that the only way of increasing the size of the data sample is to expand the time period of interest. Another possible consequence of the fact that the observations are measured over time is that the variables are time-dependent, which implies that the errors are not uncorrelated but, to the contrary, dependent on errors from earlier time periods.

Moreover, some time periods include regressor and response values that highly deviate from the rest of the data sample, for instance the financial crisis that lasted between 2007-2009. The potential implications of varying regressors and response values for different time periods is that the importance of the individual regressors is dependent on the current economic situation. A possible method for combating the issue of outliers that arise as a consequence of an economic crisis would be to create two different models, where one model would be applicable for normal economic conditions and the other for deep recessions. Although this approach was considered, it was deemed infeasible due to the limited number of observations. For instance, the time period 2007-2008 would only include 24 observations.

Regarding how well the model actually addresses the research questions, it is furthermore relevant to discuss to which extent the indexes actually represent the two different sectors of interest. To begin with, the indexes only include companies that are large enough to be listed on the stock exchange, implying that the proportion of the sectors corresponding to non-listed companies is not considered at all in the model. Further on, listed companies are in general more susceptible to the opinions of the general public. Consequently, it can be argued that changes in economic conditions, for example expansions and recessions, have a larger effect on listed companies than non-listed companies. When interpreting the results of the model, it is worth acknowledging these two limitations of the current data as it directly affects the ability to answer the research questions.

### 6.3 Future Research

To begin with, it is suggested that future research considers what influence time has on the model and if any time-dependence can be identified. The time aspect was unfortunately omitted from this thesis given the scope of the project. However, it would be appropriate to perform a time series analysis when performing further research within the subject. One possible method for analyzing multivariate time series and providing forecasts is to perform a vector autoregression. As part of the procedure, vector autoregression also facilitates the determination of causalities between the variables, which is why vector autoregression is especially suitable for this subject. [42, p. 385-386]
Another suggestion for future research would be to attain different models corresponding to different economic conditions with the aim of determining if the importance of the regressors is dependent on economic conditions. As previously mentioned, this aspect was excluded from this thesis due to lack of data. A viable option would, however, be to more thoroughly examine relevant databases in order to attain observations that are beyond the current time interval. Future research should also consider to analyze other industries in order to further investigate whether industries are subject to different macroeconomic variables.

7 Conclusion

As stated in the introduction of this thesis, the aim is to determine macroeconomic variables that are important in explaining the stock return variations in two different industries and furthermore the share of the variation explained by these selected variables. As a result, the following two research questions are to be answered:

- Which macroeconomic variables are important in explaining stock return fluctuations within banking and industrial goods and services?
- To what extent do macroeconomic variables explain stock return fluctuations within banking and industrial goods and services?

To begin with, the results of the analysis illustrate that money supply, repo rate, inflation, GDP, SEK/EUR exchange rate, consumer expectations, and spot price of oil are significant in explaining the stock return fluctuations within industrial goods and services. With regard to the banking industry, the analysis indicates that money supply, inflation, GDP, SEK/USD exchange rate, SEK/EUR exchange rate, and spot price of oil are significant in explaining the stock return fluctuations. Given that relevant literature and research within the field have been reviewed and acted as foundation for selection of the macroeconomic variables to be included in the full model, the first part of the aim of the thesis is considered fulfilled.

Secondly, the results indicate that money supply, repo rate, inflation, GDP, SEK/EUR exchange rate, consumer expectations, and spot price of oil explain 96.65% of the variation in stock return within industrial goods and services and that money supply, inflation, GDP, SEK/USD exchange rate, SEK/EUR exchange rate, and spot price of oil explain 81.76% of the variation in stock return within the banking industry. As stated in the discussion, this result is not considered fully reasonable since logically relevant variables, such as company and industry specific information, are not included in the model. Consequently, it is acknowledged that this part of the research question cannot fully be answered.

In conclusion, it has been shown that macroeconomic variables are indeed important in explaining stock return fluctuations within banking and industrial goods and services. Furthermore, the results indicate small differences in which macroeconomic variables are important within each industry, whilst the majority of the variables are common. Moreover, it was not possible to determine to what extent the selected macroeconomic variables explain the variation in stock return within the two industries. However, the high values of the coefficient of determination can possibly be attributed to the phenomenon that the regressor variables illustrate trends over time. Since this conclusion is moreover supported by economic theory, it is recommended that future research includes time series analysis.
Further on, as previously mentioned, a limitation of the model is that it does not determine causality, and consequently it is recommended that future research incorporates the causality aspect in order to draw more insightful conclusions. It is also likely that the results would differ for other industries and therefore future research should consider performing a similar analysis where other industries are the subjects of focus, with the over-arching goal of reaching a more diversified discussion and conclusion.
8 References


9 Appendices

List of Figures

1 Partial Regression Plots - Bank index ........................................ 26
2 Partial Regression Plots - Industry Index .................................... 26
3 Standardized Residuals vs Fitted Values .................................... 27
4 Standardized Residuals vs Regressors - Bank Index ....................... 28
5 Standardized Residuals vs Regressors - Industry Index .................. 28
6 Normality Plot ........................................................................... 29
7 Analysis of Influence - Bank Index ............................................. 30
8 Analysis of Influence - Industry Index ....................................... 30
9 Analysis of Influence - COVRATIO ............................................ 31
10 Analysis of Influence - DFFITS .................................................. 31
11 Analysis of Influence - DFBETAS .............................................. 31
12 Correlation Matrix ................................................................... 32
13 Lasso Regression ....................................................................... 33
14 All Possible Regression - Bank Index ......................................... 34
15 All Possible Regression - Industry Index .................................... 35
16 Normality Plot - Final Models .................................................... 38

List of Tables

1 Summary Statistics - Bank Index ................................................ 24
2 Summary Statistics - Industry Index .......................................... 25
3 Variance Inflation Factor ........................................................... 32
4 Mean Squared Error Comparison ............................................... 34
5 Summary statistics - Final Industry Index Model .......................... 36
6 Summary statistics - Final Bank Index Model ............................... 37
7 Variance Inflation Factor - Final Industry Index Model ................ 37
8 Variance Inflation Factor - Final Bank Index Model ..................... 37