Can IPO first day returns be predicted? A multiple linear regression analysis

AMAR GALIJASEVIC
JOSEF TEGBARU
Can IPO first day returns be predicted? A multiple linear regression analysis

AMAR GALIJASEVIC
JOSEF TEBBARU
Abstract

During the last three years the Swedish stock market has showed a strong upwards movement from the lows of 2016. At the same time the IPO activity has been large and a lot of the offerings have had a positive return during the first day of trading in the market.

The goal of this study is to analyze if there is any particular IPO specific data that has a correlation with the first day return and if it can be used to predict the first day return for future IPO’s. If any regressors were shown to have any correlation with the first day return, the goal is also to find a subset of regressors with even higher predictability. Then to classify which regressors show the highest correlation with a large positive return. The method which has been used is a multiple linear regression with IPO-data from the period 2017-2018.

The results from the study imply that none of the chosen regressors show any significant correlation with the first day return. It is a complex process which might be difficult to simplify and quantify into a regression model, hence further studies are needed to draw a conclusion if there are any other qualitative factors which correlate with the first day return.
Sammanfattning

Under de senaste tre åren har den svenska aktiemarknaden visat en kraftigt upptagnaende rörelse från de låga nivåerna 2016. Samtidigt har det varit hög IPO-aktivitet, där många noteringar har haft en positiv avkastning under den första handelsdagen.


Studiens resultat visar att ingen av de valda regressorerna har någon signifikant korrelation med avkastningen under första handelsdagen. Börsintroduktioner är komplicerade processer som kan vara svåra att förenkla och kvantifiera i en regressionsmodell, men ytterligare studier behövs för att dra en slutsats om det finns andra kvalitativa faktorer som kan förklara utvecklingen under första handelsdagen.
Acknowledgement

First and foremost we would like to thank SVD Börsplus, specifically Marcus Modin at IPO-guiden for providing us with data and being helpful by answering questions throughout the project. Secondly we want to thank both Jimmy Olsson and Per Jörgen Säve-Söderbergh, at the department of Mathematics, for their contribution by answering questions when such arose during the study.
# Contents

1 Introduction ................................................. 1
   1.1 Background ........................................... 1
   1.2 Aim of the study .................................... 2
   1.3 Research questions ................................. 2
   1.4 Scope ............................................... 3
   1.5 Previous studies ................................. 3

2 Theoretical framework ..................................... 5
   2.1 Economical aspect of the study ................. 5
      2.1.1 Initial public offering .................... 5
      2.1.2 The stock market ........................... 6
   2.2 Mathematical aspect of the study ............ 8
      2.2.1 Assumptions of the linear regression model .............. 8
      2.2.2 Multiple linear regression analysis ............. 8
      2.2.3 Indicator variables ....................... 9
      2.2.4 Residual analysis ....................... 9
      2.2.5 QQ-plot ................................... 12
      2.2.6 Plots of adjusted variables .............. 12
      2.2.7 Diagnostics, leverage and influential observations ....... 12
      2.2.8 Multicollinearity ....................... 14
      2.2.9 Methods for dealing with multicollinearity ........... 15

3 Methodology .................................................. 17
   3.1 Regressor variables ............................. 17
      3.1.1 IPO listing place ............................ 17
      3.1.2 Percentage of offering pre-subscribed ............ 18
      3.1.3 Gross offering ............................ 18
      3.1.4 Offering share price ........................ 18
      3.1.5 Market capitalization ...................... 18
      3.1.6 Amount newly issued equities ............... 19
      3.1.7 Advisor ..................................... 19
      3.1.8 Change in VSTOXX 30 days pre-IPO .......... 19
3.2 Data gathering ......................................................... 19
  3.2.1 Table of regressor names ...................................... 20

4 Results ........................................................................ 21
  4.1 Model 1 .................................................................. 21
    4.1.1 Residual analysis ............................................... 22
    4.1.2 Residuals versus fitted ....................................... 23
    4.1.3 Normal QQ-plot ............................................... 24
    4.1.4 Added Variables plots ....................................... 25
    4.1.5 Cook’s Distance and CovRatio ............................... 26
    4.1.6 Multicollinearity - Variance Inflation Factors ........... 27
    4.1.7 Selection of variables ........................................ 27
  4.2 Model 2 .................................................................. 29
    4.2.1 Residual analysis ............................................... 29
    4.2.2 Residuals versus fitted, Scale-Location and QQ-plot . 30
    4.2.3 Added Variables plots ....................................... 31
    4.2.4 Cook’s Distance and CovRatio ............................... 32

5 Discussion .................................................................... 33
  5.1 Information asymmetry ............................................ 33
  5.2 Signalling theory ...................................................... 34
  5.3 Information cascades ............................................... 34
  5.4 Media coverage ...................................................... 35

6 Limitations ................................................................... 36

7 Conclusion .................................................................... 37

8 Further studies .......................................................... 39
  8.1 Performance indicators ............................................ 39
  8.2 Twitter mentions ..................................................... 40
  8.3 Regression trees ...................................................... 40
1 Introduction

1.1 Background

Every year on the largest Swedish stock exchanges (NASDAQ Stockholm, First North, NGM MTF and Spotlight) there are dozens of initial public offerings (IPO) and during times of strong economic growth there has even been over a hundred of IPO’s per year. An American study shows that the first day return of IPO’s in Sweden between 1960-2011 in general has been over +25%, which means that companies usually are undervalued by the underwriter of the IPO.[1, pg. 821]

The greatest reason for why a company does an IPO is to raise equity capital and increase liquidity of the company’s stocks. An initial public offering is just as it sounds the first time the public gains access to the company’s stocks and it can be done either through a primary or a secondary offering. Both types of offerings demand that the price of the IPO is neither too high nor too low in order to raise much capital. An initial price which is too high might lead to the IPO not being fully subscribed and a price too low results in the shareholders not raising as much capital as they could have.

Making a prediction of future stock price is very difficult and there are full-time analysts who are constantly doing their best at valuing stocks, but no one is completely correct every time. The firm undergoing the IPO is not valued by the market, au contraire the company is valued by an underwriter. As the company is valued by an underwriter who sets the offering price of the IPO, it is of interest to analyze historical data in order to try to find any correlation with the first day return.

By doing a regression analysis on key parameters tied to an IPO, covered more in detail in coming sections which and why, it might be possible to find a correlation between some of the parameters and the return of first trading day of an IPO. The plan is to investigate how different company characteristics effect the market response of the IPO in terms of positive or negative first day return. Reaching a conclusion would enable for private investors to predict the first day return of coming IPO’s and by that hopefully get a high return. Since even a slight improvement of the predictive capability of such a model would be of importance for both the
company doing the offering, institutional investors and private investors, this study is of interest and is of high relevance.

1.2 Aim of the study

The aim of this study is to evaluate if IPO first day return can be predicted and whether it will be positive or negative by evaluating IPO characteristics. This will be done by doing a multiple linear regression analysis on historical IPO data. The regressors will be a set of different indicators describing the size and performance of the company e.g. price per share, percentage of amount secured and amount owned by key-investor.

One can presume that the first-day return of an IPO will depend on the numerical sizes of these factors. This just by the fact that most private investors will look at some of these factors before making the decision of investing or not. However, the main focus is to evaluate if, and in that case which, of these factors have a more significant effect on the first-day return. There may be factors that effect the way underwriters value the company in comparison to the valuation of the market’s of the company, which is where a private investor could find a high return.

1.3 Research questions

Considering the aim of the study, the research questions that will be evaluated are:

- Is there a set of regressors that have significant effect on the first-day return of IPO’s?
- Is there any subset of regressors that improves the model predictability?
- In order to increase the possibility of positive first-day returns of IPO’s, what regressors are most important for private investors?
1.4 Scope

The study will only consider initial public offerings on NASDAQ Stockholm, First North, NGM MTF and Spotlight during the time period 1 January 2017 to 31 December 2018. The reason behind limiting the study to these specific stock exchanges is firstly because the market is fairly well known by the authors. Secondly market structures are not the same globally, which means having no geographical constraints can lead to a bad result. Furthermore the time limit was set due to data in that time frame being accessible and as market structures change throughout time this time frame was seen as highly relevant. The limitation is not seen as an obstacle due to the fact that various IPOs have been conducted during those two years which means there is enough data to create a model.

Due to the fact that there are many influential factors in each IPO it is not possible to include them all as regressors in the study. By reason of restriction, a number of company-specific data, described more thoroughly in Section 3, were used as regressors.

1.5 Previous studies

Initial public offerings have been a common phenomena for as long as stock exchanges have existed, and issuing stocks most likely even longer than that. As a result, the subject is widely studied across the globe.

Over the years the subject of underpricing in IPO’s has been studied, and is mentioned by Berk DeMarzo in 2014 [1]. A study on the subject of underpricing in the Swedish market was done by Berggren[2] where he examined the level of underpricing in Sweden. The study showed that underpricing existed during the period of 2010-2016 which is highly relevant for this study. Berggren did not mainly focus on finding a model to predict the first day return after IPOs, but he did study if ”tech-IPOs” had a larger average first day return than firms in other sectors. Tech-firms had a high average return but at the same time a great variance in the return which suggests potential difficulties in valuing the firms. A lot of the IPOs during 2017-2018 are also tech-companies [3] meaning it can be good to remember that they tend to be difficult to value and have a high variance in the return.
Furthermore in 2010 Lowry, Officer and Schwert did a scientific study on the variability of IPO Initial returns[4]. The report state that the monthly volatility of IPO initial returns is substantial, fluctuates dramatically over time and is considerably larger during "hot" IPO markets. The reason explained is that the complexity of the pricing problem is related to both firm-specific and market-wide factors (e.g. when the market may be "hot"), and that this complexity limits underwriters’ ability to accurately value IPO’s. Firm-specific volatility are mainly caused by the fact that some firms are more difficult to value than others - which in turn may depend on information asymmetry and corresponding industry. Since this study indicates that the initial return of IPO’s depends on a specific set of factors, it is of great interest to evaluate if some factors show larger importance than others.
2 Theoretical framework

2.1 Economical aspect of the study

2.1.1 Initial public offering

Initial public offering (IPO), sometimes referred to as stock market launch, is a type of public offering in which shares of a company are sold to both institutional investors and retail investors. To note is that an IPO is underwritten by one or more investment banks, who also arrange for the shares to be listed on one or more stock exchanges.

Initial public offering is the first time the public gains access to the company’s stocks and it can be done either through a primary or a secondary offering. In a primary offering the company issues new shares to raise new capital, but in a secondary offering already existing shares are sold to the public which means that the company is not raising capital but the owners of the company shares are. Secondary offerings are usually done as a part of existing shareholders exit strategy. Either way the aim is not to undervalue the company as that leads to less capital being raised. At the same time there are investors who invest in IPO’s in order to make profit by hoping the stock price would increase during the first day of the IPO.[1, pg. 813]

As earlier described an underwriter is in charge of the IPO, in the larger offerings there is even a group of underwriters. There is usually one lead underwriter which gives most of the advice and arranges for the group of underwriters, called the syndicate, to arrange and sell the issue.

There is a lot of legal filings needed to be filed prior to an IPO which the underwriters help the company with and a part of this is what is called the preliminary prospectus, or the red herring. The prospectus circulates to possible investors prior to the IPO and contains information about the company and the IPO. [1, pg. 816]

Another very important part of the IPO, and this study, is the company valuation. Valuing a company is a time demanding task and there are several valuation methods. Most underwriters value future cash flows and compute them to a present value, or they value similar companies and previous similar IPOs. Since these meth-
ods usually give different answers most underwriters combine these methods to find the most reasonable price. The value of the company which the underwriters find reasonable might not be what the market is ready to pay, which is why the underwriters arrange for a road show after they have established a price range of the valuation. The road show is, much like it sounds, when senior managers of the company and the lead underwriter travel around the country, and sometimes even globally. They promote the company and explain their reasoning behind the valuation to institutional investors as pension funds and mutual funds. [1, pg. 817]

After the road show, the potential customers inform the underwriter of how many shares they are willing to purchase and at what share price. The underwriter then does a book building process where they sum up all the demand and adjust the price until it is not likely that the issue will fail.[1, pg. 818]

2.1.2 The stock market

A stock market is the aggregation of buyers and sellers of stocks (also called shares), that represent ownership claims on businesses. Basically it is a loose network of economic transactions, i.e. not a physical facility. Shares on the stock market may include securities that are listed on a public stock exchange, but also securities that are traded privately. When hearing the term ‘stock market’ it is common to refer to the securities that are listed on the public stock exchanges as Nasdaq, Dow jones or the New York Stock Exchange. These are infrastructure platforms that facilitates the trading of those equity securities, or stocks. However, securities that are traded privately include shares of private companies which are sold to investors through, for example, equity crowdfunding platforms. Stock exchanges also list other security types such as corporate bonds and convertible bonds.

Stock price movement is caused by the change of supply and demand. If it tends to be more buyers than sellers of the stock, the demand is greater than supply which causes the price to go up. But if it tends to be more sellers than buyers of the stock, the supply is greater than the demand which causes the price to drop. Obviously, the difficulty in predicting stock price movement is not due to the understanding of supply and demand but instead in comprehending what makes people like or dislike a specific stock. This comes down to figuring out what news or information that
is positive or negative for the company. Since the price per share is the current reflection of the aggregated valuation of the company by all stock owners, estimating future stock price is commonly known as being very difficult, and even nearly impossible. [5]
2.2 Mathematical aspect of the study

Unless otherwise explicitly stated, all mathematical theory described in section 2.2 is taken from Montogomery D. C., Peck E. A. Vining G. G. (2012).

2.2.1 Assumptions of the linear regression model

There are major assumptions that are taken in the study of regression analysis. These are shown below:

1. The relationship between response variable $y$ and regressor variable $x_j$ is approximately linear.
2. The error term $\varepsilon$ has zero mean and constant variance $\sigma^2$.
3. Errors are uncorrelated.
4. Errors are normally distributed.

Heteroscedasticity is easily discovered by examining if the variance of the modelling errors vary - if not, there is no heteroscedasticity. The modelling errors are then considered uncorrelated. [7, pg. 238-243][8, pg. 214-221]

One should always consider the validity of these assumptions to be doubtful and therefore always check the adequacy of the model. Residual analysis is a very useful method for diagnosing violations of the basic regression assumptions, it will be further explained later on in the study.

2.2.2 Multiple linear regression analysis

A regression analysis is used in this project to investigate if there is a relationship between the predictor variables and the return on first day after an IPO. As there are multiple predictor variables this is a multiple linear regression analysis. The model is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

Every parameter $\beta_j$ represents the expected change in response $y$ per unit change in $x_j$ when all of the other regressors $x_i (i \neq j)$ are constant. There are $k$ number of
predictor variables.

The true relationship between the $y$- and the $x$-variables is unknown, but the linear regression model is an adequate approximation of the true unknown function over certain ranges of the regressor variables.

An estimation of the model will be done using the function \texttt{lm} in the program \texttt{R}. But theoretically, the parameters in the linear regression model can be estimated by the least square method regardless of the form of the distribution for the errors.

Since there are multiple regressors and parameters $\beta$ as well as $n > k$ observations the model is in vector form:

$$y = X\beta + \varepsilon$$

with the parameters being estimated by $\beta = (X^\top X)^{-1}X^\top y$

\subsection*{2.2.3 Indicator variables}

In contrast to all of the other variables we have which are quantitative, i.e. they have a well-defined scale of measurement, the indicator variables (also called dummy variables) are qualitative variables. They have no natural scale of measurement, so a set of levels are assigned to them. The indicator variables used in this study are IPO listing place and advisor.

\subsection*{2.2.4 Residual analysis}

There are different types of residuals and the following are some of the most common for multiple regression analysis. They form an efficient tool for detecting both leverage and influence. By using many different kinds of residual analysis, there is a higher chance of detecting model inadequacy.

\textbf{Non-scaled residuals}

The definition of a residual is $e_i = y_i - \hat{y}_i$ where $y_i$ is an observation and $\hat{y}_i$ is the corresponding fitted value of the model. The residuals are seen as the deviation between the data and the fit, which also makes plotting the residuals an effective method of detecting violation of model assumptions. The residuals have zero mean,
\( \mathbb{E}(e) = 0 \) and their approximate average variance is estimated by

\[
\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n - p} = \frac{\sum_{i=1}^{n} e_i^2}{n - p} = \frac{SS_{Res}}{n - p} = MS_{Res}
\]

### Standardized residuals

The standardized residual is a measure of the strength of the difference between observed values and estimated values. Determining the standardized residuals is a way of detecting possible outliers in the regression model. It is estimated by

\[
d_i = \frac{e_i}{\sqrt{MS_{Res}}}
\]

and a value of \( d_i < 3 \) is an indication of a possible outlier.

### Studentized residuals

When calculating the standardized residuals one is using \( MS_{Res} \) to approximate the average variance of residuals. In the calculation of the studentized residuals the residual scaling is improved by dividing \( e_i \) with the exact standard deviation of the \( i \)th residual.

If the form of the model is correct the studentized residual will have a constant variance \( \text{Var}(r_i) = 1 \) regardless of the location of \( x_i \). In many cases, particularly for large data sets, the variance of the residuals stabilizes which means there is very little difference in information between the studentized and the standardized residuals. However, the risk is that points with large residuals and large \( h_{ii} \) may be influential on the least squares fit so it is preferred to examine the studentized residual.

The studentized residual is calculated by:

\[
r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}, \text{ where } i = 1, 2, ..., n \text{ and where } h_{ii} \text{ is the diagonal elements in the hat-matrix } H = X(X^T X)^{-1}X^T
\]

### PRESS Residuals

Another approach in order to find outliers or extreme values is by examining the quantity of residuals computed from \( y_i - \hat{y}_{(i)} \). Variable \( \hat{y}_{(i)} \) is the fitted value of the \( i \):th response based on all observations except the \( i \):th one. The reason why the \( i \):th
observation is not included in the examination is because that if it is unusual, it will influence the regression model. Influencing the regression model will then lead to a smaller difference between the fitted value \( \hat{y}_{(i)} \) and the observed \( y_i \). The result will be a smaller residual \( e_i \) which makes it more difficult to find an outlier. Therefore, by deleting the i:th observation, \( \hat{y}_{(i)} \) will not be influenced and the chance of detecting an outlier is much greater.

If we fit the model for the \( n - 1 \) remaining observations (since the i:th observation is deleted) while calculating the predicted value \( y_i \), the prediction error is

\[
e_{(i)} = y_{(i)} - \hat{y}_i
\]

which is calculated for every observation \( i = 1, 2, ..., n \). Then the PRESS (prediction error sum of squares) is defined as

\[
\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} y_i - \hat{y}_{(i)}
\]

Fortunately, fitting all \( n \) different regressions is not needed since it is possible to calculate the PRESS from the result of a single least squares fit to all the \( n \) observations.

The i:th PRESS residual is therefore estimated by

\[
e_{(i)} = \frac{e_i}{1 - h_{ii}}, n = 1, ..., n
\]

We see that when \( h_{ii} \) (diagonal elements of the hat matrix \( H = X(X^T X)^{-1}X^T \)) is large, the PRESS Residuals will also become large which results in high influence points. It is important to remember that a larger difference between the ordinary residual and the PRESS indicates a point where the data fits well.

Variance of the i:th PRESS residual is

\[
\text{Var}[e_{(i)}] = \text{Var}\left[\frac{e_i}{1 - h_{ii}}\right] = \frac{1}{(1 - h_{ii})^2[\sigma^2(1 - h_{ii})]} = \frac{\sigma^2}{1 - h_{ii}}
\]
so that a standardized PRESS residual is
\[
\frac{e_{(i)}}{\sqrt{\text{Var}[e_{(i)}]}} = \frac{e_i}{\sqrt{\sigma^2(1 - h_{ii})}} = \frac{e_i}{\sqrt{\sigma^2(1 - h_{ii})}}
\]

### 2.2.5 QQ-plot

By plotting the quantiles (percentiles from the residual distribution) against a theoretical distribution, which is the normal distribution in the QQ-plot, it is possible to determine if the variables are normally distributed. Smaller differences from the normal distribution (the straight line) will not affect the model fit greatly, while gross differences may do so.

### 2.2.6 Plots of adjusted variables

The Added Variable plot, also known as partial residual plot or adjusted-variable plot, is used to provide information about the marginal usefulness of a variable which is then currently not in the model. The Added Variable plots consider the marginal role of each regressor \( x_j \) given the other regressors that are in the model. In the plots both the response variable \( y \) and the regressor \( x_j \) are both regressed against the other regressors in the model.

If the regressor \( x_j \) enters the model linearly then the partial regression plot should show a linear relationship, meaning the residuals will fall along a straight line with a nonzero slope. The slope of the line will be the regressor coefficient of \( x_j \) in the multiple linear regression model. If the slope is zero it indicates that there is no additional useful information in \( x_j \) for predicting \( y \).

### 2.2.7 Diagnostics, leverage and influential observations

**Cooks distance**

When computing regressions each observation does not have the same weight in determining the outcome as for example the location in \( x \) space can play an important role in determining the regression coefficients.
A leverage point has an unusual x-value and may control certain model properties. But if the y-value lies almost on the regression line passing through the rest of the points it won't have that much of an effect on the regression coefficients. It has an effect on the model summary statistics though, such as $R^2$ and the standard errors of the regression coefficients. An influence point on the other hand has both an unusual x-value and y-value. An influence point has a large effect on the regression coefficients as it "pulls" the regression model in its direction. Not all leverage points are influential.

As stated it is of interest to consider both the location of the point in x space and the response variable when measuring influence. A way to do this is by measuring Cook's distance, a measure of the square distance between the least-squares estimates based on all $n \hat{\beta}$ and the estimate which is obtained by deleting the $i$:th observation $\hat{\beta}_{(i)}$. Cook's distance can be calculated as

$$D_i = \frac{\hat{r}_i^2 \text{Var}(\hat{y}_i)}{p \text{Var}(e_i)} = \frac{\hat{r}_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}, i = 1, 2, ..., n$$

$p$ is a constant but apart from that $D_i$ is a product of the squared residual of the $i$:th studentized residual, $e_i^2$ which reflects how well the model fits the $i$:th observation $y_i$ and the other component is $\frac{h_{ii}}{1 - h_{ii}}$ which measures how far the component is from the rest of the data. Generally points with $D_i > 1$ are considered to be influential.

**Covariance ratio**

A way to measure the effect of the $i$:th observation on the precision of estimation is to calculate the covratio which is defined as

$$\text{COVRATIO}_i = \frac{|(X_{(i)}^T X_{(i)})^{-1} S_{(i)}^2|}{|(X^T X)^{-1} MS_{Res}|}$$

If $\text{COVRATIO}_i > 1$ the $i$:th observation improves the precision of the estimation while if $\text{COVRATIO}_i < 1$ the $i$:th point worsens the precision.

Further cutoff values are decided by $\text{COVRATIO}_i > 1 + 3p/n$ and $\text{COVRATIO}_i < 1 - 3p/n$ and if an observation lies outside of these values it should be considered as influential.
2.2.8 Multicollinearity

If the regressors are almost perfectly linear, the problem of multicollinearity is said to exist. This will dramatically impact the regression model since it will be misleading or inaccurate, some of the problems being:
1. Identification of the relative effects of the regressor variables
2. Estimation and/or prediction of beta-coefficients
3. Selection of an appropriate set of variables

However, when the regressors are orthogonal or almost orthogonal (since most regressors in regression model applications are not completely orthogonal), the inferences will be easy to make.

From section 2 it is known that the regression model is described as

\[ y = X\beta + \epsilon \]

where \( X_j \) contains the \( n \) levels (\( y \) is \( n \times 1 \), \( X \) is \( n \times p \)) of the \( j \):th regressor variable. Formally multicollinearity is defined as the linear dependence of the columns of \( X \).

Even if there is linear dependence in a subset of \( X \), the rank of \( X^\top X \) will be less than \( p \) and therefore \( (X^\top X)^{-1} \) will not exist. As a result, it can make the usual least-square analysis of the regression model inadequate.

It is important to understand the source of multicollinearity to be able to analyze the data correctly. Four primary sources are data collection method employed, constraints on the model, model specification and an overdefined model.[6, pg. 286]

**Variance inflation factor**

There are several techniques for detecting multicollinearity and variance inflation factor is one of them. The diagonal elements of \( C = (X^\top X)^{-1} \) can be written as \( C_{jj} = (1 - R_{jj}) \). \( R_{jj} \) is the coefficient of determination obtained when \( x_j \) is regressed on the \( p - 1 \) regressors. The variance inflation factor is defined as

\[ \text{VIF}_j = C_{jj} = (1 - R_j)^2 \]

and if \( x_j \) is nearly orthogonal to remaining regressors, \( R_{jj}^2 \) is small and \( C_{jj} \) close to unity. But if \( x_j \) is nearly linearly dependent to some subset of regressors, \( R_{jj}^2 \) is close...
to unity and $C_{jj}$ large. A value of $VIF_j > 5 - 10$ indicates multicollinearity which will result in poorly estimates of the regressor coefficients, while smaller values of $VIF_j$ indicates moderate multicollinearity.

2.2.9 Methods for dealing with multicollinearity

Variable selection
Even if variable selection does not guarantee elimination of multicollinearity, it is described as the most common corrective technique of the phenomenon. Important to remember is that building a regression model based on a subset of available regressors creates two conflicting objectives. Firstly we would like to include as many of the regressors as possible in order to have more information affecting the predicted variable $y$. At the same time, we want as few regressors as possible since the increasing number of regressors also increases the variance of $y$.

All possible regressions
A computational technique for variable selection is the algorithm all possible regressions. The algorithm requires that a fit is made (based on some suitable criterion) for every regression model including one candidate regressor, two candidate regressors and so on. It is assumed that $\beta_0$ is included in every regression model and the number of equations will be $2^k$, $k$ being number of regressors. The ”best” regression model is then selected. Since a larger number of regressors will demand exponentially more equations to compute and thus taking much longer to run, this variable selection procedure is better for models with few ($k < 30$) regressors. Two criteria used in order to evaluate each subset are adjusted $R^2$ and Bayesian Information Criterion.

$R^2$ and Adjusted $R^2$
A way to measure the adequacy of the model is to use

$$R^2 = \frac{SS_R}{SS_T}$$

which is referred to as the proportion of variation explained by the regressors $x_j$. But, the reason adjusted $R^2$ is used and not simply $R^2$ is because $R^2$ never decreases
when a regressor variable is added, regardless of the variable improving the model or not. Adjusted $R^2$ can be defined as

$$R_{Adj}^2 = 1 - \frac{SS_{Res}/(n - p)}{SS_T/(n - 1)}$$

where

$$SS_T = SS_R + SS_{Res} \iff \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

which describes the total sum of squares as a sum of the model sum of squares and the residual sum of squares.

**Bayesian information criterion**

Bayesian information criterion, also called BIC, is used for variable selection among a finite set of models. As the sample size increases, the criterion places great penalty on adding regressors. If ordinary least squares regression is used, this criterion is mathematically computed as

$$BIC_{Sch} = n \ln \frac{SS_{Res}}{n} + p \ln n$$
3 Methodology

3.1 Regressor variables

3.1.1 IPO listing place

The Scope of this study is limited to initial public offerings on the stock exchanges NASDAQ Stockholm (small-, mid- and large cap), First North, NGM MTF and Spotlight. Since these regressor variables are categorical, they will be divided into indicator variables.

Becoming a publicly traded company is often a very big and expensive step for most companies. Deciding what stock exchange to do the initial public offering on depends on what the specific stock exchange offers in terms of for example price and trade revenue. Each company has different assets and long-term goals, therefore, one single stock exchange does not fulfill the needs of each company that wants to become public.

- **NASDAQ Stockholm** is the primary stock exchange in Sweden and the listed companies are divided into the sizes Small Cap, Mid Cap, Large Cap. Small Cap is for companies with market value of less than 150 million euro, Mid Cap for companies with market value between 150 million euro and 1 billion euro while Large Cap is for companies that has a market value larger than 1 billion euro. [9]

- **First North** is a division of NASDAQ Nordic. It is a stock exchange for smaller growth companies in Europe and designed to access growth capital in order to develop and expand their businesses. More than five companies grow and transfer to the Main Market each year and there are currently more than 260 companies traded on First North.[10]

- **NGM MTF** refers to the stock exchange North Growth Market (NGM) that has a division for trading called Nordic MTF, hence the name NGM MTF. NGM MTF is a platform for small- and mid-sized growth companies not yet fully developed to do their initial public offering on the larger stock exchanges. The judicial status of the NGM MTF is not as extensive and the demands are
therefore lower for listing. Thus, investments in companies on NGM MTF is also considered as more risky. [11]

- **Spotlight** is similar to NGM MTF and First North in the sense the platform wants to make it easier and safer for growth companies to become public. [12]

### 3.1.2 Percentage of offering pre-subscribed

Before private investors can subscribe to the IPO large parts of the offering are pre-subscribed by large institutions, often just after the underwriters are done with a road show. The pre-subscription is guarantee commitment without any legal binding, but as maintaining good relationships in the business is important it is often viewed as a binding commitment.

### 3.1.3 Gross offering

Gross offering is defined as the total amount of the shares which will be available in the offering.

### 3.1.4 Offering share price

The offering share price is the share price which is offered to the public private investors, which may differ from the price which large institutions sign up for during the road show. This regressor variable has continuous values.

### 3.1.5 Market capitalization

The value used to classify the value of the companies is market cap. The market cap of a company is the total value of its shares, i.e share price multiplied by total number of shares.[1, pg. 28]
3.1.6 Amount newly issued equities

A percentage amount which shows how much of the total equity in the offering is newly issued equity, the rest is already existing equity which belongs to shareholders of the company.

3.1.7 Advisor

The underwriter that lead most initial public offerings during 2017-2018 was Carnegie Investment bank, Eminova Fund Commission and Sedermera Fund Commission. therefore, these three will be chosen as indicator variables and all of the others will be put in the category ”others”.

3.1.8 Change in VSTOXX 30 days pre-IPO

This index reflects the market expectations of volatility by measuring the square-root of the implied variance across all options of a given time to expiration. The indices are based on EURO STOXX 50 real-time options prices.[13]

By measuring the change in VSTOXX in the time gap which we believe a lot of investors are likely to sign up for the IPO (30 days pre-IPO), the point is to include the short-term uncertainty of current market. Higher volatility, or change in volatility, means that a security’s value might dramatically change over a short period in either direction. This information might therefore correlate with the first-day return of an IPO. [14]

3.2 Data gathering

The response variable in the model is presented in decimal form, i.e. if the first-day return is 10% then $y$ will have a value of 1.01.

Data of all the regressor variables, excluding VSTOXX, was collected from the 'Börsplus IPO-guide’, which is a free service in the daily newspaper Svenska Dagbladet[Öörs]. This service compile data from all initial public offerings made on all Swedish stock
exchanges from the beginning of 2017 until today. The data from the site was gathered from an original excel-file compiled by Marcus Modin, an employee at SVD, in order to be used in this study. The data of regressor variable VSTOXX was downloaded from the Reuters database.

3.2.1 Table of regressor names

<table>
<thead>
<tr>
<th>Methodology &amp; Results</th>
<th>Added Variables plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of offering pre-subscribed</td>
<td>SakradAndelAvErbjudandet</td>
</tr>
<tr>
<td>Gross offering</td>
<td>ErbjudandeBrutto</td>
</tr>
<tr>
<td>Market capitalization</td>
<td>BorsvaardeEfterNotering</td>
</tr>
<tr>
<td>Amount newly issued</td>
<td>Nyemession</td>
</tr>
<tr>
<td>Offering share price</td>
<td>Noteringkskurs</td>
</tr>
<tr>
<td>Change in VSTOXX 30 days pre-IPO</td>
<td>EUROVolDiff</td>
</tr>
<tr>
<td>Nasdaq Large Cap</td>
<td>ListaNasdaq Large</td>
</tr>
<tr>
<td>Nasdaq Mid Cap</td>
<td>ListaNasdaq Mid</td>
</tr>
<tr>
<td>Nasdaq Small Cap</td>
<td>ListaNasdaq Small</td>
</tr>
<tr>
<td>NGM MTF</td>
<td>ListaNGM MTF</td>
</tr>
<tr>
<td>Spotlight</td>
<td>ListaSpotlight</td>
</tr>
<tr>
<td>Eminova Fund Commission</td>
<td>HuvudradgivareEminova FK</td>
</tr>
<tr>
<td>Sedermera Fund Commission</td>
<td>HuvudradgivareSedermera FK</td>
</tr>
<tr>
<td>Carnegie Investment Bank</td>
<td>HuvudradgivareCarnegie</td>
</tr>
</tbody>
</table>

Figure 1: The corresponding names of the regressor variables in the Added Variables plot.

The regressors are named differently in the Added Variables plot. Hence, Figure 1 describes what regressor names these correspond to in comparison to the original regressor names in section 3.1.
4 Results

4.1 Model 1

$R_{Adj}^2 \approx 0.07$

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Estimate (Beta)</th>
<th>Std. Error</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.188e-02</td>
<td>8.565e-02</td>
<td>-0.372</td>
<td>0.710279</td>
</tr>
<tr>
<td>Percent of offering pre-subscribed</td>
<td>1.320e-01</td>
<td>1.010e-01</td>
<td>1.307</td>
<td>0.193169</td>
</tr>
<tr>
<td>Gross offering</td>
<td>5.364e-11</td>
<td>2.076e-10</td>
<td>-0.258</td>
<td>0.796419</td>
</tr>
<tr>
<td>Market capitalization</td>
<td>5.553e-12</td>
<td>7.571e-11</td>
<td>0.073</td>
<td>0.941633</td>
</tr>
<tr>
<td>Amount Newly Issued</td>
<td>2.170e-11</td>
<td>1.717e-10</td>
<td>0.126</td>
<td>0.899811</td>
</tr>
<tr>
<td>Offering share price</td>
<td>1.074e-03</td>
<td>1.462e-03</td>
<td>0.753</td>
<td>0.452495</td>
</tr>
<tr>
<td>Change in VSTOXX 30 days pre-IPO</td>
<td>2.747e-01</td>
<td>2.871e-01</td>
<td>0.957</td>
<td>0.340122</td>
</tr>
<tr>
<td>Nasdaq Large Cap</td>
<td>2.687e-01</td>
<td>7.640e-01</td>
<td>0.352</td>
<td>0.725546</td>
</tr>
<tr>
<td>Nasdaq Mid Cap</td>
<td>1.093e-01</td>
<td>1.900e-01</td>
<td>0.575</td>
<td>0.566117</td>
</tr>
<tr>
<td>Nasdaq Small Cap</td>
<td>3.160e-02</td>
<td>1.376e-01</td>
<td>0.230</td>
<td>0.818733</td>
</tr>
<tr>
<td>NGM MTF</td>
<td>-3.242e-01</td>
<td>8.639e-02</td>
<td>-3.753</td>
<td>0.000251</td>
</tr>
<tr>
<td>Spotlight</td>
<td>3.663e-02</td>
<td>6.996e-02</td>
<td>0.524</td>
<td>0.601351</td>
</tr>
<tr>
<td>Carnegie Investment Bank</td>
<td>-2.357e-02</td>
<td>1.305e-01</td>
<td>-0.181</td>
<td>0.856900</td>
</tr>
<tr>
<td>Eminova Fund Commission</td>
<td>-1.891e-01</td>
<td>2.418e-01</td>
<td>-0.782</td>
<td>0.435460</td>
</tr>
<tr>
<td>Sedermara Fund Commission</td>
<td>5.564e-01</td>
<td>3.393e-01</td>
<td>1.640</td>
<td>0.103223</td>
</tr>
</tbody>
</table>

Figure 2: Initial full model using the function summary() in program R.

Figure 2 shows the initial full model of all regressors. Percent of offering pre-subscribed, Change in VSTOXX 30 days pre-IPO, Nasdaq Large Cap and Sedermara Fund Commission where the ones with highest positive effect on first day return. The stock exchange NGM MTF is the only one that shows a significantly large negative value.
4.1.1 Residual analysis

Studentized residuals

A plot of the studentized residuals against the corresponding fitted values is an easy way of detecting model inadequacies. If the residual values can be contained in a horizontal band it is an indication of negligible model defects.[6, pg. 139] The plot seems to fulfill this criteria except one single observation in particular that stands out. That particular observation is from the IPO of Stenocare which had a very high first day return in comparison to the rest.

PRESS Residuals

The plotted PRESS Residuals against the corresponding fitted values shows, in combination with the studentized residuals and figure 4, that the mean of all errors are near zero. Analyzing PRESS Residuals is good for detecting single observations that are seriously influencing the overall fit of the model. Observation 147, Stenocare, is an example of such a phenomena. This is consistent with the studentized residuals, so the plot of PRESS residuals only made it more clear.

Assumption 2: the errors $\epsilon$ has zero mean, is fulfilled.
4.1.2 Residuals versus fitted

The impression from the residuals versus fitted plot is that it is a bit fluctuating, but stable according to the almost-straight red line. Moreover, there is no specific pattern between fitted values and residuals which means the error terms are uncorrelated and have constant variance. Scale-Location and residuals versus fitted show no sign of heteroscedasticity.

**Assumption 2:** the variance of the error terms being constant, is fulfilled.

**Assumption 3:** the error terms being uncorrelated, is fulfilled.
4.1.3 Normal QQ-plot

The QQ-plot shows that the standardized residuals, errors, follows the normal distribution line in the middle but deviates from it towards the right. Observation 147 Stenocare being the point that deviates the most.

**Assumption 4** The errors may therefore be considered as near normally distributed.
4.1.4 Added Variables plots

Unfortunately most Added Variables plots of the regressors near-zero slopes, indicating that they have almost no effect on the first-day return. The regressor variables *SakradAndelAvErbjudandet*, *ListaNGM MTF*, *EUROVolDi* and *HuvudradgivareSedermera* show a slight linear relationship with a nonzero slope. However, the reason that the latter has a nonzero slope is because of one single observation.
Looking at all of the observations for this regressor, they are the opposite of equally spread out across the line and the removal of this single observation would therefore significantly affect the slope.

**Assumption 1**: the relationship between response variable $y$ and regressor variables $SakradAndelAvErbjudandet$, $ListaNGM MTF$, $EUROVolDiff$ is approximately linear.

### 4.1.5 Cook’s Distance and CovRatio

![Figure 8: Cook’s Distance and Covratio of initial full model](image)

The plot of Cook’s distance does not show any influence since all points fulfill $D_i > 1$. However, the plot of Covratio states that the majority of the points, all that are below 1, degrade the precision of the model. [6, pg. 215-219]
### 4.1.6 Multicollinearity - Variance Inflation Factors

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of offering pre-subscribed</td>
<td>1.164484</td>
</tr>
<tr>
<td>Gross offering</td>
<td>20.132038</td>
</tr>
<tr>
<td>Market capitalization</td>
<td>18.502174</td>
</tr>
<tr>
<td>Amount Newly Issued</td>
<td>4.178758</td>
</tr>
<tr>
<td>Offering share price</td>
<td>1.623756</td>
</tr>
<tr>
<td>Change in VSTOXX 30 days pre-IPO</td>
<td>1.101049</td>
</tr>
<tr>
<td>List</td>
<td>36.025219</td>
</tr>
<tr>
<td>Main advisor</td>
<td>2.063002</td>
</tr>
</tbody>
</table>

Figure 9: Variance Inflation Factors of initial full model

Some of the regressors imply serious problems with multicollinearity because of values larger than 10 [2, p.118]. List, Percent of offering pre-subscribed and Market capitalization have by far the largest VIF values. The rest of the regressors have VIF values below 5-10 which is low enough to not affect the model severely.

### 4.1.7 Selection of variables

The Bayesian Information Criterion (BIC) is a method for model selection among a finite set of models. It penalizes the complexity of the model where complexity refers to the numbers of parameters (regressor variables). The BIC value for each parameter compared, as defined in section 2.2.8, is calculated and shown in Figure 10. The best model is the one that provides the minimum BIC value, but as can be
seen in figure 10 it only contains one regressor. As the process of an IPO can not be simplified down to only one regressor, we chose to incorporate three regressors. Taking into account both the BIC and adjusted $R^2$, the subset-model will contain SakradAndelAvErbyudandet, ListaNGM MTF, EUROVolDiff. This will reduce the complexity of the model while not removing too many regressor variables.[16]
4.2 Model 2

$R^2_{Adj} = 0.1078$. The reduced regression model is showing a slightly higher R-adjusted value than the full model.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Estimate(Beta)</th>
<th>Std. Error</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.03470</td>
<td>0.05700</td>
<td>0.609</td>
<td>0.544</td>
</tr>
<tr>
<td>Percent of offering pre-subscribed</td>
<td>0.08793</td>
<td>0.09176</td>
<td>0.958</td>
<td>0.339</td>
</tr>
<tr>
<td>Change in VSTOXX 30 days pre-IPO</td>
<td>0.23679</td>
<td>0.27133</td>
<td>0.873</td>
<td>0.384</td>
</tr>
<tr>
<td>NGM MTF</td>
<td>-0.36213</td>
<td>0.07607</td>
<td>-4.761</td>
<td>4.27e-06</td>
</tr>
</tbody>
</table>

Figure 11: Coefficients of the Reduced Model

However, the main purpose of this study was to evaluate if first-day returns can be predicted. Since this final reduced model is only able to explain 10.78% of the variance present in the data, the correlation between regressors and first-day return seems to be very low.

*Change in VSTOXX 30 days pre-IPO* is as expected still the largest positive coefficient and *NGM MTF* is the only negative one.

4.2.1 Residual analysis

![Studentized residuals and PRESS Residuals of Reduced Model](image)

Figure 12: Studentized residuals and PRESS Residuals of Reduced Model

```r
> mean(fit$residuals)
[1] -2.392966e-17
```

Figure 13: Mean of errors of Reduced Model
When comparing the studentized residuals plot in Model 1 to the one in Model 2, there are no major differences. The same applies to the PRESS residuals and the mean of the errors.

**Assumption 2:** errors having zero mean, is fulfilled.

### 4.2.2 Residuals versus fitted, Scale-Location and QQ-plot

![Residuals versus fitted, Scale-Location and QQ-plot](image)

**Residuals versus fitted and Scale-Location**

The residuals versus fitted plot slightly differs from Model 1. Although, the residuals do not increase or decrease with the fitted values in any pattern which states the fact that the variance is still constant. Scale-Location plot does not show any sign of heteroscedasticity.

**Assumption 2:** the variance of the error terms being constant, is fulfilled.
Assumption 3: the error terms being uncorrelated, is fulfilled.

QQ-plot
The normal QQ-plot shows no difference in comparison to Model 1, hence the model may be considered as almost normal distributed.
Assumption 4: fulfilled.

4.2.3 Added Variables plots

These Added Variable plots were obtained by reducing the full Model 1 into Model 2. The plots for the regressor variables *SakradAndelAvErbjudandet* and *EUROVolDiff* are almost flat, meaning that there is very little linear correlation between them and the response variable, first-day return. The fact that the slight linearity showing in Model 2 decreased even further in Model 1, indicates that the linearity is very low.

Regressor variable *ListaNGM MTF* still shows a non-zero line in Model 2, yet it can be seen most of the observations are separated into two big areas. The points
are not equally distributed along the blue line. Still listings on *NGM MTS* would indicate negative return during 2017-2018.

**Assumption 1:** Some linearity can be observed.

### 4.2.4 Cook’s Distance and CovRatio

![Figure 16: Cook’s Distance and Covratio for Reduced Model](image)

Covratio seems to be showing roughly as many points below 1 as in Model 1, these points degrading the precision of the regression model. The plot also shows lower overall Covratio-values which worsens the precision of the model. Cook’s distance shows higher values as well, however no values above 1.
5 Discussion

The results of both the Full Model and the Reduced Model are shown in section 4 and, unfortunately, neither one of them show very good predictability. The Reduced Model had the best Adjusted $R^2$ of the two, with the low value of 0.1078. Section 5 discusses factors that could affect the first day return, but were not included because of the difficulty of incorporating them in a qualitative and mathematical study of this type.

5.1 Information asymmetry

As described in the introduction and the economical framework, it is not an easy task for the underwriters to set a 'correct' offering price for an IPO. It is a question of valuing the company correct while also being able to predict what the market is willing to pay for the shares. This since the market does not have the insight in the company as the owners do, which means they might not be willing to pay as much as the sellers want. This phenomena can be described as information asymmetry and is one of the major reasons underwriters underprice the offering.[17]

The first-day return of an IPO is in general over +25% [1, pg. 821] in Sweden which may be an indication that underwriters underprice the IPO intentionally. There is scientific literature suggesting this and one of the most-well known theories to explain it is proposed by Kevin Rock’s 'the winner’s curse'. This theory arises from the problem that to win an auction, the winner has to outbid the others which leads to the winner of the auction ending up with the highest valuation of the object. In the case of IPO’s, the winner’s curse is relevant since the aftermarket performance tends to be greater when there is great significant demand of the issued stock, which implies that each bidder gets fewer stocks. At the same time, larger allotment of shares will be available if the demand is low which will then worsen the aftermarket performance. According to Rock, underwriters therefore compensate the uninformed investors for the asymmetric information by setting low offering prices which explains the nature of new issues being underpriced.[17]

Furthermore, the IPO’s made during 2017-2018 on the Swedish stock exchanges
have mostly been within industries known for having high research and development expenditures. Biotech, medtech and gaming being the most common ones. According to Chang and Su, information asymmetries are notably high within these industries which may be caused by the fact that their valuation methods differ from the traditional, established, companies[18]. However, they reach the conclusion that RD investments induce information asymmetries which raises the level of underpricing.[19] This might explain the difference between the first-day returns, and thereby also the significant variance in the model. An improvement of the regression model would be to add one regressor variable describing the industry of the observation, which would confirm or neglect the claim.

5.2 Signalling theory

A further theory behind why IPO’s are initially underpriced is the Signalling theory. It is assumed that the best information about the prospect is within the firm itself. Wishing to signal to investors about the superior prospects, firms do an offering at a low price and quantity. By doing this the firm ”leave money on the table” meaning they do not earn as much equity as they could have by having a higher initial offering price. This type of signalling is a type of credibility signal, or equilibrium signal of firm quality, that the firm is good, as only a good firm can be able to recoup this type of loss after it is realized. The goal by doing this is to have a strong development of the stock price shortly after the offering and then the firm can do a new issue of stocks to recoup some of the money left on the table. [20, pg. 303-306]

The strong initial development of stock price is hopefully also seen as a positive signal by the rest of the market which can attract further investors.

5.3 Information cascades

Another theory firmly connected with the last part of the signalling theory, about attention from media, is the theory about information cascades. Even though all investors have equal information and are able to build their opinion, many still become uncertain in themselves because of the investment-decisions of the larger crowd. This is a typical case of herd behaviour, following the behaviour of the big
mass. Having a price which is a bit too high might result in the IPO not being fully subscribed, scaring further investors off. [21, pg. 12-13]

Both the theory of signalling and information cascade are difficult to quantify. Capturing their effect in the regression model of this study would be of great importance, but unfortunately very difficult. The fact that information cascades was excluded in the model, while still possessing an important role in first-day returns, could explain the low predictability of the model.

5.4 Media coverage

Just as the factors described in Discussion, media coverage is a factor which is hard to quantify in an easy and reliable way. What the media writes before an IPO and especially in which way they address the specific firm in the IPO affects how attractive it becomes for investors.[22] This matter was not addressed in the model of this study, hence possibly an additional explanation to the the low predictability of the model.

Allen and Faulhaber discuss the importance of media even further in their study about the Signalling theory. Allen and Faulhaber discuss the possible gain of publicity a company obtain by the initial underpricing. As previously described the firm hopes for a strong initial development by underpricing. There are often articles in financial newspapers about recent IPOs with a strong development. Such publicity by credible financial newspapers gives a strong positive signal to the rest of the market about the firm.[20, pg. 306] A further thought which builds upon what Allen and Faulhaber have written is that an initial low price can lead to great publicity as financial news sources might write articles about how good the offering price is. Meaning it will attract a large mass of buyers and the price will most likely have a positive development shortly after the IPO as not everyone will get the initial subscription they wished for.
6 Limitations

One of the most significant limitation of this study was the fact that observations only was gathered from the time period 2017-2018. By gathering data from earlier time periods as well, more nuanced and solid results could have been presented. However, the model includes enough observations to determine that the correlations between response- and regressor variables are low.

Another major limitation are the chosen variables in the study. There is a possibility that other IPO specific data e.g. the available period for investors to sign the IPO or industries (including performance indicators and Twitter mentions) would have shown better correlation with the first-day return. Especially industries, since the danish company Stenocare influenced the regression model greatly - reason being it is the first cannabis-company being listed on a Swedish stock exchange[23]. A completely new industry with high demand increased the offering price 213% the first day which is incredibly high in relation to any other companies first-day returns. However, having one regressor being industries was not possible to do because of the data collected from SVD Börspuls IPO-guiden did not provide that type of data. Unfortunately, alternatives as Bloomberg Terminal or similar databases were too complicated to download data from.
7 Conclusion

The aim of this study was to evaluate if it is possible to predict the first day return after IPO’s by choosing a set of regressors with possible impact and doing a multiple linear regression with historical IPO-data. This by firstly doing a regression model with the complete set of regressors, and secondly with a reduced set of regressors that increased the predictability of the model. Lastly the study evaluates what regressors had significant impact on positive first-day return, but also other factors with potential impact on the return that were not included in the model.

In the first part of the study all of the chosen regressors were included in the model and it showed a very weak linearity, a surprisingly low $R^2_{adj} \approx 0.07$. The largest positive coefficients were Percent of offering pre-subscribed, Change in VSTOXX 30 days pre-IPO and Sedermera Fund Commission. In the Added Variable plot for Sedermera FK it is clear that one single observation influences the graph heavily, hence the coefficient is not reliable for a general model. The largest negative coefficient was NGM MTF followed by Eminova Fund Commission. Accordingly, to get a large positive first day return one wants a large percentage of secured amount, a high volatility of Eurostoxx, Sedermera FK to be the main advisor and should stay away from offerings on NGM MTF with Eminova FK as the main advisor.

Further a Reduced Model including three regressors were chosen after validation of the initial full model using Bayesian Information Criterion and Adjusted $R^2$. The regressors are now only Percent of offering pre-subscribed, Change in VSTOXX 30 days pre-IPO and NGM MTF. A combination of both positive and negative coefficients. The model showed slightly better adjusted $R^2_{adj}(= 0.1078)$, but the value still needs to be higher in order to rely on the model for unknown data. The Added Variable plots still showed near-zero slopes which indicates the regressors have very little correlation with the first day return.

The conclusion of the study, with the two models done, is that there is very little correlation between the chosen variables and the first day return of IPO’s. An investor should not use this particular model with these regressors for prediction of first day returns.

Additional studies need to be made with other regressors, where key performance
indicators and Twitter mentions are two proposed candidates of regressors. Furthermore topics as information asymmetry, the winner’s curse and the Signalling theory have been discussed as they, amongst others, might be reasons for why IPO’s in general are underpriced and are difficult to incorporate in a mathematical regression study.
8 Further studies

8.1 Performance indicators

The income statement of a company provides very useful information when estimating the profitability of a firm’s business and how that relates to the value of a firm’s shares. Most common ratio that analysts and investors use to gauge the market value of the firm is the firm’s price-earnings ratio \( P/E \):[1, pg. 30-33]

\[
P/E = \frac{\text{MarketCapitalization}}{\text{NetIncome}}
\]

However, an IPO valuation is very dependent on the company’s future growth projections. Meaning that a significant part of value creation is growth and the primary motive is therefore to raise more capital in order to fund further growth. Hence, qualitative components are not always an effective tool in valuating an IPO, but estimating Return On Equity (ROE) to find any indication of the firm’s capability to find investment opportunities could be very useful[24, pg. 30-33][1]. This is estimated by:

\[
ROE = \frac{\text{NetIncome}}{\text{BookValueofEquity}}
\]

Since performance indicators similar to these are the most common tool in valuation of a company, to include them in the regression analysis would give an aspect to the study of great importance.

Unfortunately the gathering of this particular data was too complicated because of two main factors; firstly it would demand a manual search of through every 168 prospectus of the IPO’s in order to find the correct financial data. Secondly, most growth companies fluctuates heavily just quarters a part because of their financial instability which may cause major errors in the calculations of the ratios. With this being said, including key performance indicators is a recommendation for further study but should be correctly formatted and gathered from the corresponding advisory if possible.
8.2 Twitter mentions

One way to measure the impact of information cascades for the first-day returns of IPO’s would be by counting tweets for each company doing an IPO a specified time period before the IPO, and then put it as a regressor variable. Twitter is one of the most popular micro-blogging social platforms on the web. An approximate 22% of all internet-users in Sweden has sometime used the service and 6% uses it daily[25]. It is therefore natural to assume that tweets and retweets about a specific company is very common when it is pursuing an initial public offering. According to information cascades, this positive or negative buzz will then affect a larger mass to either sign the IPO or not. Depending on whether the larger mass have positive, negative or neutral comments, information cascades states the effect of twitter mentions would represent the beta-estimate accordingly.

However, this was not done in this study due to the fact that Twitter only offer API for tweet-counts for the last 7 days. In order to get access of the application programming interface (API) for the full archive, Twitter charges a high fee[26].

8.3 Regression trees

In this study it has been evaluated if it is possible to predict the IPO first day return with a multiple linear regression. Another method to use could be a tree-based algorithm which might have given a better or clearer result. A regression tree is a hierarchical display of series of questions that relate to each unit in the sample. Answering all these questions will then give the answer to what the first day return will be.

This is usually represented as an upside down tree where the root is at top, a series of branches is connecting the nodes and the leaves are at the bottom. At each node a question about the variables is asked and depending on the answer, one of the other branches is chosen. That means that the order of which the questions are asked in is of high importance. Usually questions which maximize the node purity are at the top, where node purity is improved when the node variability in the response data at the node is minimized. If the response is discrete then high purity would be a low amount of classes or categories. On the other hand if the response is continuous
then measures of variability as standard deviation or mean square error should be made as small as possible to maximize node purity. [6, pg. 524]

Another approach is to make this into a classification problem. Simplifying the problem into two classes, positive or negative return, instead of the full percentage scale as in this study. The process in this case would be similar as the one for the regression tree. These types of algorithms are usually called CART (Classification And Regression Trees). [6, pg. 524]

Other than regression trees it could be interesting to use a neural network which is often used to model high-dimensional, non-linear data. [6, pg. 526]
References


https://www.nyemissioner.se/foretag/planerad-noteringar/sok [2019-02-04]


https://www.investopedia.com/articles/basics/04/100804.asp


http://www.ngm.se/om-ngm/[2019-02-17]

https://spotlightstockmarket.com/om-spotlight [2019-02-17]


[16] Richards, D. Maximum Likelihood Estimation and the Bayesian Information Criterion


https://sedermera.se/sv/uppslag/2018-10/stenocare [2019-03-26]


[26] Twitter (Developer).