Stockholms universitet

Predicting Tomorrow's Direction of the Swedish Stock Market

Philippe Goldmann

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:31
Matematisk statistik
December 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm
University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2016:31**
http://www.math.su.se

Predicting Tomorrow's Direction of the Swedish Stock Market

Philippe Goldmann*

December 2016

Abstract

This thesis examines the possibility to predict tomorrow's stock market direction with the information we have today. It uses existing market data to create new variables which in turn function as explanatory variables for predicting tomorrow's direction. Different models are tested during different market conditions and the tests have been made on official Swedish stock exchange data. This study examines the predictive performance in financial markets with logistic regression during an 8-year period.

The study finds models with poor discrimation of predictive abilities on the swedish stock index OMXS30. Yet, data indicate that some predictive ability exists during normal and stressed market conditions. Some models achieve to keep a majority of predictions correct in all tested settings.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: philippe_goldmann@hotmail.com. Supervisor: Jan-Olov Persson & Gustav Alfelt.

Abstract

This thesis examines the possibility to predict tomorrow’s stock market direction with the information we have today. It uses existing market data to create new variables which in turn function as explanatory variables for predicting tomorrow’s direction. Different models are tested during different market conditions and the tests have been made on official Swedish stock exchange data. This study examines the predictive performance in financial markets with logistic regression during an 8-year period.

The study finds models with poor discrimation of predictive abilities on the swedish stock index OMXS30. Yet, data indicate that some predictive ability exists during normal and stressed market conditions. Some models achieve to keep a majority of predictions correct in all tested settings.

Acknowledgments:

I would like to thank Jan-Olov Persson and Gustav Alfelt for guidance in computations and modelling. I would like to thank Riksbanken, CBOE and NasdaqOMXNordic for the easy accumulation of data. I would further like to thank my father Stefan Goldmann for teaching me mathematics at a very young age. I would also say thanks for the support from family and friends.

Table of contents

Chapter one: “Predicting Tomorrow’s Direction of the Swedish Stock Market”

1. Introduction, Background, Variables and Problem Discussion, Motivation of the study, Purpose and aim of the study, Research Questions, Structure & Contribution

1. Introduction

The introduction chapter begins with a background and with a problem discussion. This chapter also contains the motivation, aim and purpose of the study. This is followed by research questions, the structure of the thesis and general contribution.

1.1 Background

The structure of the modern stock market has its roots in 1531, Antwerp Belgium. Brokers and moneylenders used it as a meeting point dealing in business, government and personal debt issues. During this day and age, no shares exchanged hands. Instead, stakeholders used promissory notes. (Sowani, 2013)

Today, the stock market is a vast entity dealing with financial information, prices, news and many aspects of modern day life. For each year technologies and innovations are introduced making the market itself increasingly complex. Therefore, the factors behind the movements of the stock markets have changed, is changing and will continue to change in the future.

From an investor’s point of view, this could turn problematic. If an investor were to glance at historical prices, it could encourage the belief that the future can be determined by the past. However, with constant renewals of the development of markets, one must ask, will any mathematical model prevail predicting the future with consistency?

Figure 1.1 Dow Jones price history year 1890 - 2010

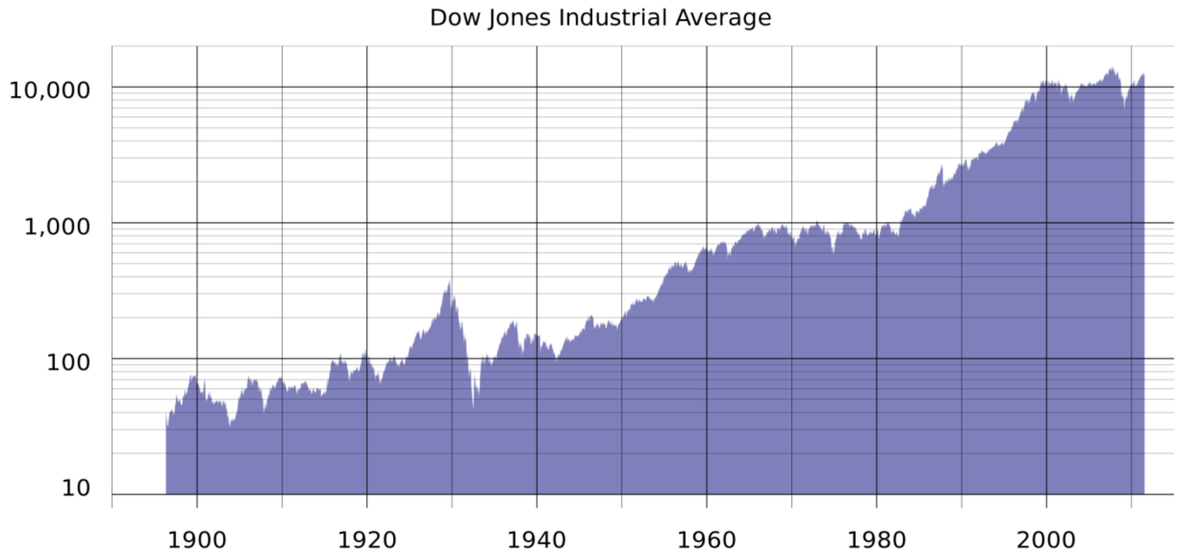


Figure source: http://www.djaverages.com/#indexData

As we might obtain from figure 1.1 of the Dow Jones price history, there seems to be a general up-going trend during these 120 years observed. Thus, investors might assume the market always goes up in the long-term perspective. Yet we know, looking at daily returns of the stock market, there are many factors which effects the direction of the daily returns. Sometimes the stock market goes up and sometimes it goes down. When viewing the daily returns, we might not have the 120-year perspective in the back of our mind. In fact, most of the author's friends care mainly about a very short-term trend of the stock market. This is also relevant to the finance industry when modelling short-term business opportunities and short-term risks. Therefore, we might be more interested in what happens tomorrow, given the information we have today.

1.2 Variables and Problem Discussion

Percentage change of OMXS30

First and foremost, we need to define the percentage change in order to discuss our variables:

Let X_{t-1} and X_t denote random indices of financial data at time *t-1* and *t*. Then the daily percentage change $R_{t-1,t}(X)$ from *t-1* to *t* is:

$$R_{t-1,t}(X) = \frac{X_t}{X_{t-1}} - 1$$

Thus, we introduce our first explanatory variable: The percentage change of OMXS30

Let P_{t-1}^{OMXC} and P_t^{OMXC} denote the closing prices of OMXS30 at time *t-1* and *t*. Then the percentage change of OMXS30 $R_{t-1,t}(P^{OMXC})$ from *t-1* to *t* is:

$$R_{t-1,t}(P^{OMXC}) = \frac{P_t^{OMXC}}{P_{t-1}^{OMXC}} - 1$$

The percentage change of a price is often called a *return* of that price. Though, as we will see in the following variable discussion, we need to compute percentage changes of variables that are not denoted as prices. Therefore, it might be found unconventional to speak about returns of volumes, rates etc.. So, we will stick to the notation of *percentage changes* instead of *returns* for the time being.

What the percentage change of OMXS30 is measuring is the direction (whether there has been a rise or fall in OMXS30) and the magnitude of that direction (if it fell, how much did it fell in percent?).

Response variable

By "direction" hereafter we define the percentage change being either positive or negative. If the stock index of OMXS30 rises from day *t* to day *t+1* the direction will be "up". If the index falls from day *t* to day *t+1,* then the direction will be "down". The direction of either "up" or "down" is regarded as a binary response which in statistics, may be approached with logistic regression. The response variable asserts numerical values giving a binary indicator for us to use the price direction in computations. So, what we would like to model is a response whether tomorrow will be up or down. This is done by constructing Y_t which gives a variable of 0 if tomorrow is "down" and 1 if tomorrow is "up":

$$Y_t = \begin{cases} 0 \; if \; R_{t,t+1}(P^{OMXC}) < 0 \\ 1 \; if \; R_{t,t+1}(P^{OMXC}) \geq 0 \end{cases}$$

Volume of OMXS30

One of the most common indicator combined with financial prices is the volume of the instrument. The volume is the number of shares or contracts traded in a security or an entire market during a given time. (Investopedia, 2016) The theory of the volume's effect on stock prices is that if the direction of the price and volume is the same, the general trend of the stock price will continue. In this study, the volume of OMXS30 is defined as the second explanatory variable. (Bourquin, 2007)

The volume itself turn problematic due to the stochastic properties of buying/selling on stock markets. Which day the market participants are performing financial transactions and for what amount of capital, might be regarded as a stochastic variable. Here the magnitude factor is important as increased or decreased selling or buying is believed to influence stock movements. This problem is addressed with a percentage change of the volume. This renders a variable that takes the day-to-day magnitude effect into account:

Let V_{t-1}^{OMX} and V_{t}^{OMX} denote the volumes of OMXS30 at time *t-1* and *t*. Then the percentage change of the volume of OMXS30 $R_{t-1,t}(V^{OMX})$ from *t-1* to *t* is:

$$R_{t-1,t}(V^{OMX}) = \frac{V_t^{OMX}}{V_{t-1}^{OMX}} - 1$$

Daily Price Range

Another explanatory variable that is derived of the price of OMXS30 is the daily price range *R*. The daily price range is defined as the difference between the highest and the lowest price during a trading day. (Investopedia, 2016)

$$Range(P_t^{OMX}) = P_t^{OMX\_high} - P_t^{OMX\_low}$$

The daily price ranges also pose problems, since range varies as a function of price. This means that higher prices will have higher ranges and lower prices have lower ranges. This is addressed through a new variable: price range in percent, giving a price range in terms of percent. This variable is computed by taking the range and dividing it with the daily closing price of OMXS30.

$$Range\ in\ percent(P_t^{OMX}) = \frac{P_t^{OMX\_high} - P_t^{OMX\_low}}{P_t^{OMXC}}$$

Vix Index

Yet another variable of the price is volatility, which can be measured from different perspectives. The definition of historical volatility is the degree of variation of a price over time as measured by the standard deviation of returns. The most common use is defining volatility over daily returns. In this paper the use of historical volatility is scarce, instead, we use the VIX index as an explanatory variable. VIX index or VIX is computed on implied volatility. Implied means that it's a volatility function that is derived from a market traded derivative estimating future volatility. This index is said to measure the market's expectation of volatility over the next 30-day period. (Avellaneda, 1995) (Brenner & Fand Galai, 1989)

Stibor Rate

Another variable to the price is interest rate. Interest rate is the amount of interest due per period as a proportion of the amount lent. In this thesis, the representation of interest rate is the STIBOR (Stockholm Interbank Offered Rate) which is a reference rate that shows the average interest at which several active banks are willing to lend to one another. (NasdaqOMXNordic, 2016) The choice of STIBOR as an explanatory variable instead of any given interest rate from the government is because mostly banks trade within the stock market. Therefore, the STIBOR-rate represent the actual price of money when banks are investing. (Sveriges Riksbank, 2012)

Lags and Dummy

To create lags first and foremost we compute percentage changes of Vix, Stibor, Volume and Daily Price Range in percent. When this has been done, we find 9 explanatory variables where 5 of them are percentage changes. These 5 changes have a certain "actuality". This implies that which price the VIX had the 10th of October in the year of 1995, is hardly relevant today. However, perhaps a volatility spike 3 days ago, has an impact today, or a volume spike 5 days ago, has an impact tomorrow. To measure this "actuality", we define 5 time lagged variables on each of these 5 percentage changes to capture the latest movements in each variable. These lags are called distributed lags and work to improve our model. They measure the interplay of day-to-day effects which can be hidden otherwise. This gives us 34 explanatory variables working to predict one response (if tomorrow is "up" or "down"). One dummy variable has been added to these 34 and it is however today is an "up" or "down" day. Thus, we have a rich model with 35 variables to help us model probabilities of tomorrow's outcome:

Table 1.2.1 All explanatory variables

Percentage change of OMXS30
Volume
Percentage change of Volume
Daily Price Range in percent
Percentage change of Daily Price Range in percent
Vix
Percentage change of Vix
Stibor
Percentage change of Stibor
Percentage change of OMXS30 lag 1-5
Percentage change of Volume lag 1-5
Percentage change of Daily Price Range lag 1-5
Percentage change of Vix lag 1-5
Percentage change of Stibor lag 1-5
Dummy variable (if today is “up”)

These changes and their lags resemble binary classifiers or dummies. Why? Well, if we note a variable decreasing from day *t-1* to *t,* the percentage change of that variable is < 0. Say volume is 100 day 1 and decreases to 80 day 2. Then the percentage change of the volume is $\frac{80}{100} - 1 = -0{,}2\ or - 20\%$. Inversely, if the volume increases from 80 to 100, the same variable is 0,25 or +25%. So, the percentage changes have a categorical property of a binary indicator because if we have a rise of a variable we obtain a positive change. And in the same manner we obtain a negative change if the variable decreases. Thus, applying a dummy variable of 1 if we have a positive change, and 0 otherwise, does not add any information to the model. We already know which direction the variables are traveling since we see if their changes are positive or negative. Thus, we are settling with the 35 variables stated so far.

1.3 Motivation of the Study

The study aims to be a quantitative research about logistic regression models in mathematical statistics targeting applied finance. The approach is deductive due to the problem formulation: We construct our model, then we try to establish as good predictions as possible given different market conditions. Then we try to link these results with general theory.

The motivation of the study is to raise the question if it's possible at all to predict the direction of OMXS30 and with what accuracy? The topic is of interest for the entire financial sector and claims for/against the idea of being able to predict the stock market is a controversial topic.

This study might as well invoke a spark of interest for students in mathematical statistics which have an interest in finance. It might contribute to the general understanding of stock market uncertainty and the ability of predicting stock indices.

1.4 Purpose and Aim of the Study

This study aims to reach a general understanding regarding prediction of stock market outcomes. Comparing models conclude in different questions that is answered throughout this thesis. The purpose is to find whether logistic regression might be a tool at hand for the given time interval to assess predictive properties on financial time series.

1.5 Research Question

Does logistic regression provide good prediction results in normal and stressed market conditions given different model selections?

1.6 Structure

The second chapter is a theoretical background to logistic regression and statistical theory regarding it. The third chapter is describing the method and models, datasets and criticism. The fourth chapter contains the analysis and model outputs. The fifth chapter presents results, limitations as well as ideas for future research. Thereafter a conclusion follows with references and appendices.

1.7 Contribution

The author wishes to widen general knowledge of logistic regression as a case study for all type of pass/fail questions as well as to introduce a systematic approach of how problems in finance can be quantified with mathematical statistics.

Chapter two: “Predicting Tomorrow’s Direction of the Swedish Stock Market”

2. Multiple logistic regression, Classification Accuracy, Misclassification Measurement of Multicollinearity

2.1 Multiple Logistic regression

As introduced in *Applied Logistic Regression, Second Edition. By* David W. Hosmer and Stanley Lemeshow:

If we were to consider a collection of p independent variables in the vector

$$\boldsymbol{x}' = (x_1, x_2, \dots, x_p)$$

and letting the conditional probability that the outcome is present as

$$P(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x})$$

we can express the multiple logistic regression as

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p ,$$

where the logistic function is

$$\pi(\boldsymbol{x}) = \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{1}{1 + e^{-g(x)}} .$$

If a nominal scaled variable has *k* possible outcomes, then *k-1* design variables will be needed. When we introduce design variables (in this study only one dummy variable is used, however being a type of design variable) suppose that the j^{th} independent variable x_j has k_j levels. The $k_j - 1$ design variables will be defined as D_{jl} and the coefficients for these will be β_{jl} , $l = 1, 2, \dots, k_j - 1$. Then the logistic function for a model with p variables and with the j^{th} variable being discrete:

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + \beta_p x_p$$

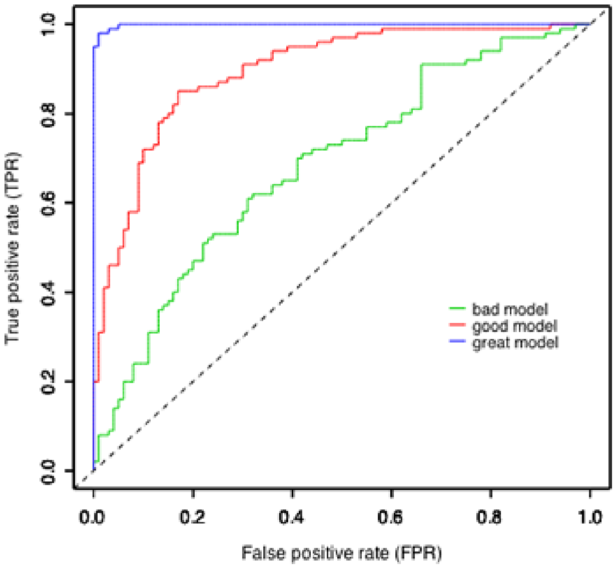
2.2 Classification Accuracy

Receiver Operating Characteristic or ROC is a complete description of classification accuracy. The ROC curve was first developed from signal detection theory. It shows how the receiver operates the existence of signal in presence of noise. It plots the probability of detecting true signals vs false signals for a range of thresholds.

The true positive rate is known as sensitivity, which measures the proportion of positives that are correctly identified. This can be interpreted as the percentage of “up” days predicted which are “up”. Whereas the false positive rate is known as the fall-out and may be calculated as 1 – specificity. Specificity measures the proportion of negatives that are identified as negatives. When observing false positive rate this can be interpreted as the percentage of “up” days predicted that would be correctly identified as “down” days. The plotting of various thresholds displays for which values optimal sensitivity and specificity is achieved.

When analysing the ROC-curve one would wish for the values of the true positive rate to be as far from the false positive rate as possible. A ROC-graph typically looks like a signal as resembled in figure 2.4.1:

Figure 2.4.1 Example of a ROC-curve



In statistical studies a popular measure is the Area Under Curve or AUC. The AUC calculates area under the ROC curve which provides a measure of discrimination. The AUC can vary between 0 and 1 where AUC = 0.5 suggests no discrimination (guessing works as well).
If $0.5 < AUC < 0.7$ then this is considered poor discrimination.
If $0.7 \leq AUC < 0.8$ then this is considered acceptable discrimination.
If $0.8 \leq AUC < 0.9$ then this is considered excellent discrimination.
If $AUC \geq 0.9$ then this is considered outstanding discrimination.

Figure Source:

http://jxieeducation.com/2016-09-27/Beautiful-Properties-Of-The-ROC-Curve/

(www.mathworks.com, 2016)

(Hosmer & Lemeshow, 2000)

2.3 Misclassification Measurement of Multicollinearity

The variable inflation factor or VIF is a measurement in R's *car* package which measures multicollinearity. The VIF is defined as:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination for a model where the $i{:}th$ explanatory variable is fit against all other variables in the model. Multicollinearity is when a predictor in a model can be replaced or approximated with a linear combination of the other explanatory variables in the model. This is mainly used for variable removement when modelling. The test gives numerical values where we reject a variable if we find a VIF of 5 or greater (in R this is a generalized VIF, adjusted for logistic regression models).

Another step of validating the model is by performing a Durbin Watson test, which tests if the errors are correlated. Here a Durbin Watson statistic is given with an associated p-value for us to reject or approve the null hypothesis and if we find symptoms of autocorrelation in residuals. This statistic is defined as below:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

where e_t is the residual at observation t.

There are as well some empirical verifications of discovering multicollinearity. One can produce *crPlots* in R to try to figure this out with analytical skills (see appendix 6.2). Coefficients of different models tend to become sensitive and behave erratically under the influence of multicollinearity.

(The Minitab Blog, 2016)

(San Diego State University, 2013)

Chapter three: “Predicting Tomorrow’s Direction of the Swedish Stock Market”

3. Method and Model, Model Selection, Data Sample Analysis and Distributions, Criticism

3.1 Method and Model

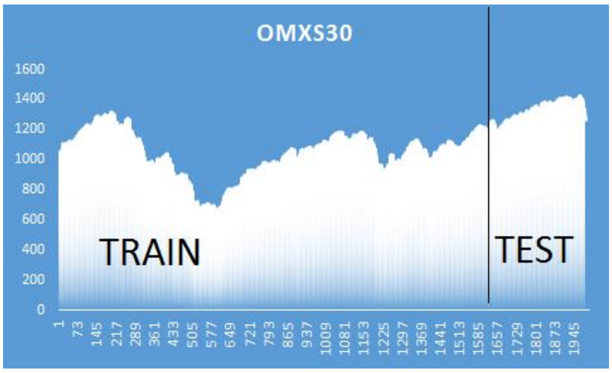
The method at hand is binary multiple logistic regression. The data for OMXS30 is recovered from the Nasdaq’s Nordic site. (NasdaqOMXNordic, 2016) There we obtain high-, low- and closing prices as well as the volume. Stibor data is gathered from Riksbanken’s website. (Riksbanken, 2016) VIX data is obtained from the CBOE website. (CBOE, 2016)

Data is downloaded to Microsoft Excel, sorted and exported to R (statistics software) via Notepad (text software compatible with R).

Robustness tests

The dataset is divided into two samples. The first data sample is the training data and the second sample is the testing data. We train the logistic model on 80% training data and test on 20% testing data as seen in figure 3.1.1:

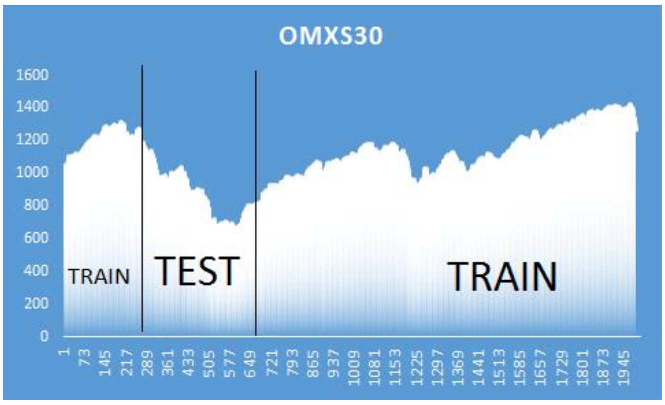
Figure 3.1.1 Division of intervals in normal market conditions



Two robustness tests are conducted in the study. In these, the datasets have the same proportions of training and testing data (80% / 20%), but the testing and training set is different for each robustness test.

The first robustness test is conducted on the financial crisis where we have a clear downward effect of the stock market index. This time interval is defined as the testing set and all other data is training data as seen in figure 3.1.2:

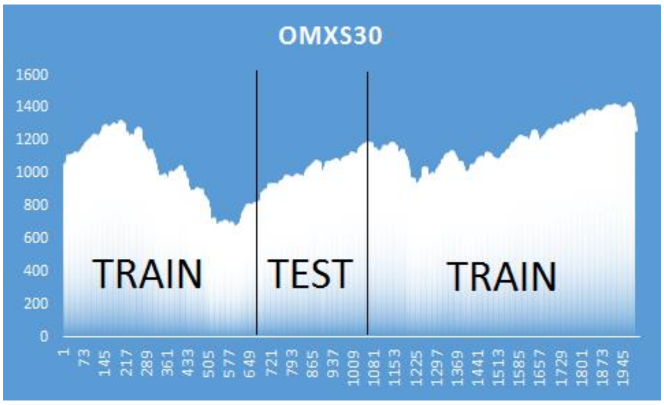
Figure 3.1.2 Division of intervals in first robustness test R1 (financial crisis)



Note that we use the both *train* datasets to predict the testing set. One might be confused by this procedure since it is not linear with time. Because tomorrow's outcome is modelled to be explained of mostly what happens today, these two training periods in figure 3.1.2 might be merged into one training period. The lags have not been manually adjusted when merging these training sets. This is due to considering that day *t* contains a number of lags which are categorized for day *t*, meaning that it would be incoherent to adjust the merged training set with new lags.

The second robustness test is conducted on the immediate period after the financial crisis where we have a clear upward effect of the stock market index. Here this interval is defined as the testing set and all other data being training data.

Figure 3.1.3 Division of intervals in second robustness test R2 (after financial crisis)



The type of testing set as seen in figure 3.1.3 contains mostly days with "up":s. This is to have transparency in robustness tests. If we were to find a model *A* which is excellent at predicting "down":s but not "up":s then this model *A* could survive R1 but not R2. Thus, having two robustness tests we test predictive abilities with both "up" and "down" days without favoring models which only possess good predictive abilities in one direction.

Data Approximation

One data approximation that must be made, is when it's an American but not a Swedish holiday. Then OMXS30 is priced but the VIX index is not. The number of missing data points of VIX correspond to 48 of 2079 days (or 2,31% of the length of the dataset). Since it is given that there are various methods of replacing data, the author found that the average of yesterday and tomorrow to be a good data replacement method.

Lags procedure

The procedure when creating lags is done by putting previous information into today's category. For instance, at any given time t we have the percentage changes of all the variables from the 5 latest days. So, at day 10 we have the percentage change of OMXS30 from day 9 to 10 ($R_{9,10}(P^{OMXC})$). Then we find our 5 lags: ($R_{8,9}(P^{OMXC})$, ($R_{7,8}(P^{OMXC})$, ($R_{6,7}(P^{OMXC})$, ($R_{5,6}(P^{OMXC})$ & ($R_{4,5}(P^{OMXC})$.

This procedure has been done with the percentage changes of all other variables (VIX, Stibor, Volume, Daily Price Range in percent). For the sake of simplifying notation, let's denote all lags as *lags*.

Model with lags

Further on, we recall from section 2.1 we found that the logistic equations were

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

and

$$\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) \; and \; \pi(\boldsymbol{x}) = \frac{1}{1 + e^{-g(x)}}.$$

Let $\boldsymbol{x}$ be a vector of all data of the variables mentioned above and let $\boldsymbol{\beta_X}$ represent the matrix which gives each lag its β-coefficient, then

$$g(\boldsymbol{x}) = \beta_0 + \; \beta_1 P_t^{OMXC} + \beta_2 Y_{t-1} + \beta_3 V_t + \beta_4 P_t^{VIXC} + \beta_5 r_t + \beta_6 R_t + \boldsymbol{\beta_X} \sum_{i=1}^{5} lags$$

where

$$P(Y_t = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-g(x)}}$$

3.2 Model Selection

Different models are tested throughout the study. The main full model contains all 35 variables and is denoted M1. The norm in our study could be to perform a backward stepwise selection of models from M1 in order to find the model with best predictive abilities. However, given the computations, R was only able to find backwards selected models filtered by AIC (Akaike Information Criterion). When tested, these models rated by AIC did not result in any good models (models that could be distinguished from random guesses). Performing backward stepwise selection by supplementary packages in R filtering by significance (p-values), the fitting algorithms for this procedure did not work.

So, the model selection has been done manually following significant variables. All robustness test models are denoted with R1 or R2 to respective model. Robustness test 1 of M1 is denoted M1R1 and robustness test 2 of M1 is denoted M1R2. M2 & M3 have descriptive qualities and M4-M8 are stepwise stripped from insignificant variables by logical reduction of variable relevance. Here are the following models and their respective structure:

M2 – All variables but no dummy, no lags and no percentage changes

omx_return + omx_vol + vix + stibor + omx_range_norm

M3 – All variables but no lags

omx_return + indicator_variable + omx_vol + vix + stibor + omx_range_norm + Treturn + Vixreturn + StiborR + Rang.n.R

M4 – All variables that are at the 5% significant level (in comparison to M1)

omx_return + indicator_variable + stibor + OMXlag1 + Tlag5 + Vixreturn + Vixlag1

M5 – All variables of M4 but without Tlag5 & Vixlag1

omx_return + indicator_variable + stibor + OMXlag1 + Vixreturn

M6 – All variables of M5 but without any lags

omx_return + indicator_variable + stibor + Vixreturn

M7 – All variables of M6 but without stibor

omx_return + indicator_variable + Vixreturn

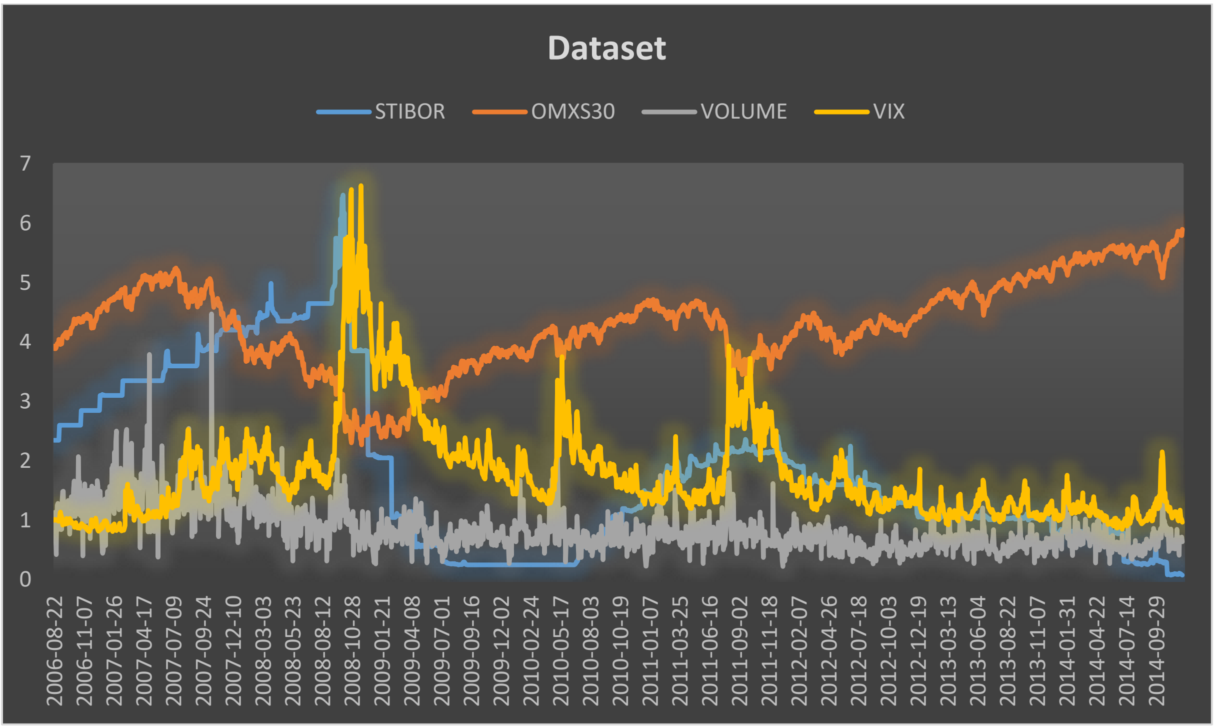
M8 – All variables of M7 but without omx_return

indicator_variable + Vixreturn

3.3 Data Sample Analysis

Data is gathered from 2006-08-30 to 2014-12-05 which contains 2079 trading days.

Figure 3.3.1 Representation of dataset

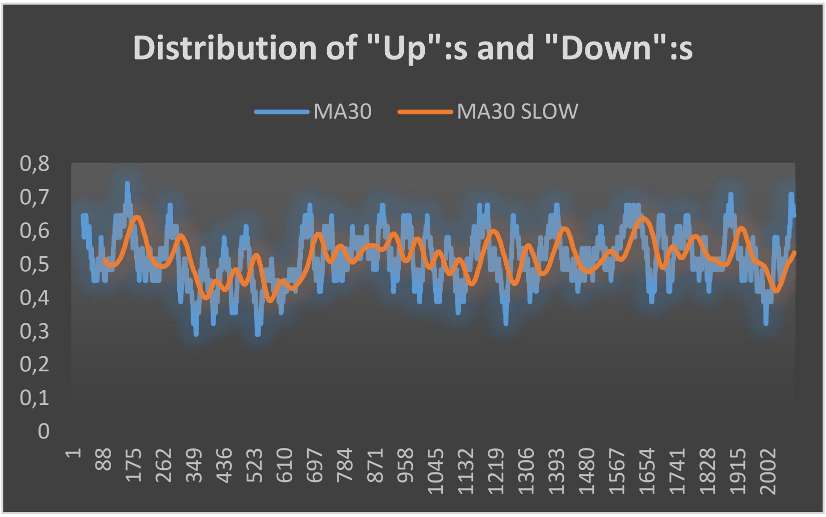


All variables in the figure 3.2.1 have been normalized to a starting value which is suitable for the graphical representation. We note that the STIBOR-rate and the VIX-index are very volatile while the OMXS30-index is much smoother. The volume graph in this figure appears as a stable signal with only a few major spikes (For a full dataset plot, see Appendices 6.1 / for full variable distribution see Appendices 6.3).

From April 2007 to February 2009 we note a downward trend in OMXS30 which was called the financial crisis of 2008. We note that STIBOR-rates depreciated sharply because of the American rescue packages to stabilize the ongoing crisis. The VIX rose sharply when the American Senate voted against rescuing one of the former biggest banks in the U.S., i.e. Lehman Brothers. Thus, this period is chosen for our robustness test type 1 (R1). The immediate period after, is chosen for our robustness test type 2 (R2). (Investopedia, 2016)

General Trend of the Response Variable

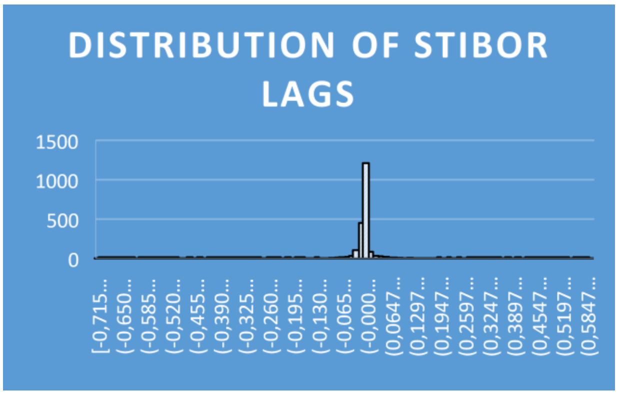
Figure 3.3.2 General Trend of Up:s vs Down:s



In figure 3.2.13 we see the rolling 30 day mean of “up”:s vs “down”:s with a slow moving average giving us an idea of the trend. “Up” is asserted with 1 and “down” is asserted with 0. What we can see is that the long-term trend (MA30) has a wavy pattern where the period of the waves differs. With this 30-day average we find that in general it tends to be in the interval of 0.7 – 0.3 and that we find a threshold of how many “down”:s or “up”:s is probable before a trend reversal.

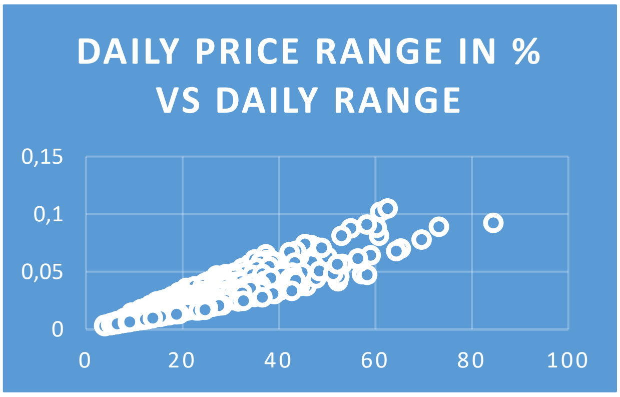
3.3 Criticism

Figure 3.3.1 Distribution of Stibor Lags



As we note in figure 3.3.1 the Stibor Lags do not distribute well in a variable sense. One might conclude that by approximating these lags by zero, it would render little to no effect on the logistic regression model. However, for the consistency of variable treatment these lags are included. (For all variable distributions please see appendix 6.3)

Figure 3.3.2 Daily Price range in percent vs Daily Range



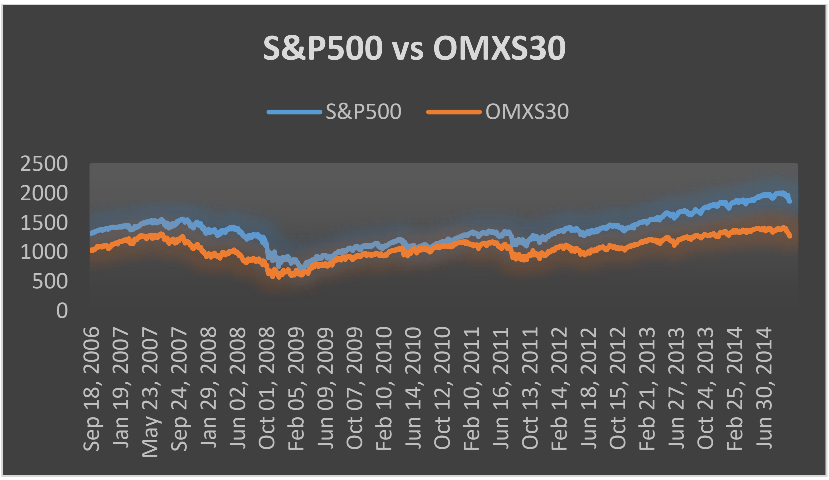
As we might conclude from figure 3.3.2 the choice of daily price ranges in percent (y-axis) is done by constructing a new variable from the ordinary daily range (x-axis). The new variable has a higher coefficient of variation by 28%. One must ask, is it necessary to introduce higher variation in the model?

One could leave the range variable in its original form and be fine with the results. Nevertheless, the reason behind computing the the variable in percent, is that deviations not measured in percent can be subject to skew the analysis. The daily range varies with the price of OMXS30 and it is obvious that the range will differ if the price of OMXS30 is 100 or 1000. Sometimes, variables are more properly measured in percent, even though they introduce higher variance.

Why do we use VIX and not the Swedish volatility index SIX?

The SIX-index is mainly data for the Swedish finance sector and the author could not find access to this data without charge. The VIX index is measured on S&P500 which OMXS30 follows closely. By using VIX as a proxy for SIX, the author has an assumption that little information is lost. This could be a wrong assumption. However, the correlation coefficient between OMXS30 and S&P500 is 0.907188 or 90,72% measured from 2006-09-18 to 2014-10-16. This indicates that the historical volatility of OMXS30 is highly correlated with S&P500. Thus, it is assumed that the pricing of implied volatility should be somewhat correlated between these two stock indices.

Figure 3.3.3 S&P500 vs OMXS30



Chapter four: “Predicting Tomorrow’s Direction of the Swedish Stock Market”

4. Analysis, Model M1, Model M8, Comparison of M8 with other models, General effect of different explanatory variables, Misclassification analysis

4.1 Analysis

ACC.P% is total accurate predictions in %, PDOWN is accurate “down” predictions in %, PUP is accurate “up” predictions in %, AUC is the ROC estimator Area Under the Curve, R1 denotes robustness test 1 (financial crisis) and R2 denotes robustness test 2 (immediate period after the financial crisis), MEAN is the mean of all values.

Table 4.1.1 Output data of M1

	ACC.P%	PDOWN	PUP	AUC
M1	0.606715	0.61111	0.604396	0.63263
M1R1	0.47482	0.42384	0.50376	0.46438
M1R2	0.48201	0.42754	0.50896	0.459
MEANM1	0.52118	0.4875	0.53904	0.51867

What we might obtain from table 4.1.1 is that the full model M1 predicts well during normal market conditions but when tested in R1 & R2 it fails in terms of keeping accurate predictions over 50%. Interestingly, in terms of predictive abilities M1 fails severely predicting “down” days but keeps a decent hit rate of predicting “up” days during robustness tests. The AUC is decent during normal conditions though we would have wished for a higher AUC in R1 & R2.

Model M2-M7 performed differently and for all full view of these models please see appendix 6.1.2.. Roughly M2 to M7 are all derived from M1, and they are simpler models trying to reduce variables and increase predictability. M2-M7 do outperform M1 in some instances, but the real contender to M1 is the most reduced model M8 following below:

Table 4.1.2 Output data of M8

	ACC.P%	PDOWN	PUP	AUC
M8	0.58513	0.56354	0.60169	0.61296
M8R1	0.57314	0.65672	0.53357	0.6275
M8R2	0.52518	0.49419	0.54694	0.56762
MEANM8	0.56115	0.57148	0.56073	0.60269

What we can see from table 4.1.2 is that M8 outperforms M1 in terms of accurate predictions in normal and stressed market conditions. This also applies when viewing the overall ability to predict up and down in all settings. The mean AUC is also better in M8. Thus, M8 is therefore a more robust model than M1 because it is more consistent within different market conditions. Modelling unknown scenarios, we wish for a model which have the prerequisites of not failing during stressed settings. Even better, M8 consists of very few explanatory variables namely 2 vs the 35 of M1.

For the overall descriptive understanding of the variables, M1 have more variable information meaning that we can generalize from it. As stated, the response variable is 1 if tomorrow is “up” and 0 if tomorrow is “down”. Given that majority of the data of the variables are positive, the sign of the coefficient describes if the explanatory variable increases or decreases the probability of the event of tomorrow being an “up” day. More on this in section 4.5.

4.2 Model M1

1662 training days

417 testing days

36 variables, total: 74 844 data points

In M1 we find the following variables to be significant:

3-star significance *** (p value less than or equal 0.001)

The return of VIX

2-star significance ** (p value less than or equal 0.01)

The Stibor rate, The first lag of OMXS30

1-star significance * (p value less than or equal 0.05)

The intercept, The return of OMXS30, Indicator Variable (if today is an “up” day), The fifth lag of Volume, The first lag of Vix

1-dot significance · (p value less than or equal 0.1)

Volume, The third lag of Volume, The second lag of normalized range

(to see all variables coefficients and p-values, please see appendix 6.4)

Table 4.2.1 Prediction output M1

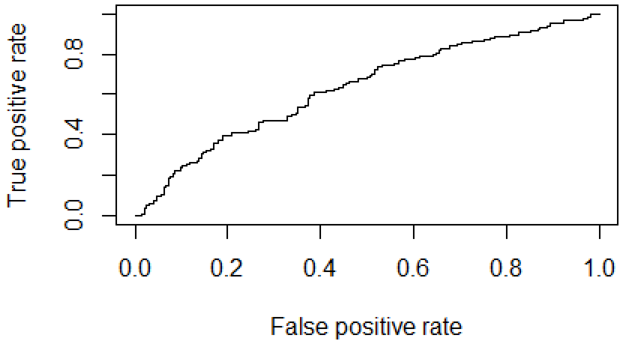
	PRED	PRED	PRED	PRED
	DOWN/DOWN	DOWN/UP	UP/UP	UP/DOWN
M1	88	56	165	108
M1R1	64	87	134	132
M1R2	59	79	142	137

Table 4.2.2 Tot. M1 pred.

PRED	
DOWN	PRED UP
144	273
151	266
138	279

As we can see in table 4.2.1 are the different predictions of both “up” and “down” predictions vs the direction the OMXS30 index has at closing time. In table 4.2.2 we note the total predictions regardless of which direction the stock index travels. What we might note is a general tendency of M1 predicting “up” days instead of “down” days.

Figure 4.2.3 ROC-curve of M1



In figure 4.2.3 we see how the ROC curve shapes for M1 and yields an AUC of 0,63. As recalled the true positive rate is considered “up” predictions which are correctly classified as “up”. The false positive rate is “up” predictions on “down” days. This means that overall M1 is finding true signals yet not in a satisfactory manner. We would have wished for M1 to categorize the predictions with better accuracy. This would imply a sharper rise of the true positive rate vs the false positive rate.

Adjusting for the AUC value of 0,5 (no discrimination) vs M1, we find a 26,52% better chance of finding true predictions with M1 compared to guessing. This also means that this model reached 26,52% of optimal performance which could determine every outcome correctly. The mean AUC in all conditions is 0,51867. Thus, implying the stressed market conditions of R1 & R2 to be equally probable as normal market conditions, M1 offers a 3,73% better chance of finding true signals compared to the value of no discrimination.

4.3 Model M8

1662 training days

417 testing days

3 variables, total 6237 data points

In M8 we find the two explanatory variables to reach 3-star significance (P-value < 0.001) namely the *indicator_variable* (if today is an "up") and the *Return of Vix* (percentage change of Vix).

Table 4.3.1 Prediction Output M8

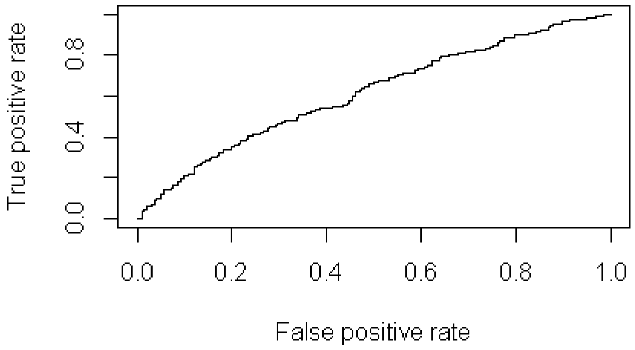
	PRED DOWN/DOWN	PRED DOWN/UP	PRED UP/UP	PRED UP/DOWN
M8	102	79	142	94
M8R1	88	46	151	132
M8R2	85	87	134	111

Table 4.3.2 Tot. M8 pred.

	PRED DOWN	PRED UP
M8	181	236
M8R1	134	283
M8R2	172	245

In table 4.3.1 we find the predictions of M8 where both correct and incorrect predictions are displayed. In table 4.3.2 we find total predictions regardless of accuracy. We can see the general tendency of predicting "up" days instead of "down" days displayed here as well.

Figure 4.3.3 Roc Curve of M8

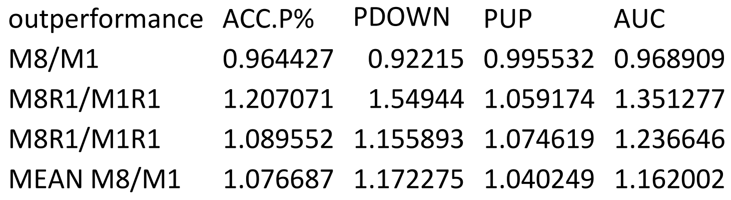


As we note in figure 4.3.3 we find a similar ROC curve to M1. The AUC is 0,6130 or 61,30% and indicate a 22,6% better chance than guessing. The mean AUC is 0,6027 implying that if R1 and R2 is equally probable M8 offers an overall 20,54% better chance of finding true predictions than the value of no discrimination.

4.4 Comparison of M8 with other models

Comparing the M1 & M8 models we compute a performance table of 4.1.1 & 4.1.2. Thus, we view the percental improvement of model M8 compared to M1:

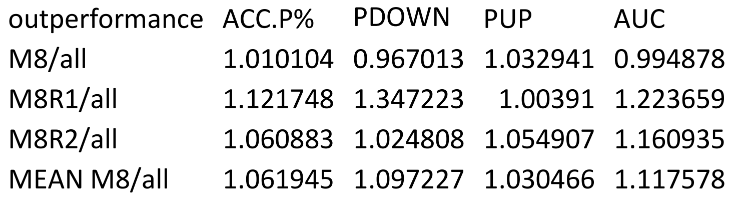
Table 4.4.1 Performance Table M8/M1



outperformance	ACC.P%	PDOWN	PUP	AUC
M8/M1	0.964427	0.92215	0.995532	0.968909
M8R1/M1R1	1.207071	1.54944	1.059174	1.351277
M8R1/M1R1	1.089552	1.155893	1.074619	1.236646
MEAN M8/M1	1.076687	1.172275	1.040249	1.162002

As we can see in table 4.4.1 the mean accurate predictions increase 7,67% with M8 compared to M1. The average ability to predict down increases with 17,23% and the ability to predict up increases with 4,02%. We find a 16,20% improved AUC. Though, M1 is stronger than M8 in normal market conditions.

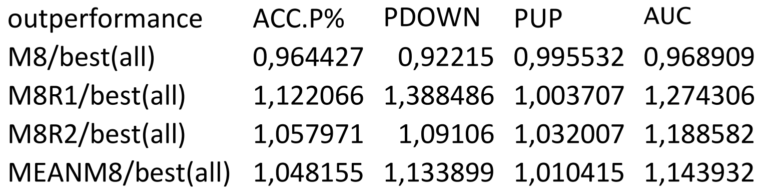
Table 4.4.2 Performance Table M8/all models



outperformance	ACC.P%	PDOWN	PUP	AUC
M8/all	1.010104	0.967013	1.032941	0.994878
M8R1/all	1.121748	1.347223	1.00391	1.223659
M8R2/all	1.060883	1.024808	1.054907	1.160935
MEAN M8/all	1.061945	1.097227	1.030466	1.117578

As we note in table 4.4.2 the overall better predictive abilities of M8 compared to all models (M1 – M7) improve. The accurate prediction increases with 6,2%, "down" predictions with 9,7% and "up" predictions with 3,0%. Still, M8 is weaker in normal market conditions with regards to predicting "down" days and with regards to the AUC.

Table 4.4.3 Performance Table M8/best values of M1, M3-M7



outperformance	ACC.P%	PDOWN	PUP	AUC
M8/best(all)	0,964427	0,92215	0,995532	0,968909
M8R1/best(all)	1,122066	1,388486	1,003707	1,274306
M8R2/best(all)	1,057971	1,09106	1,032007	1,188582
MEANM8/best(all)	1,048155	1,133899	1,010415	1,143932

As we note in table 4.4.3 we find that M8 do provide with exclusively better results when compared to the maximal values of all the other models. One might note that model M2 is excluded from the comparison. The reason for this was that model M2 behaved quite oddly and in normal market conditions M2 was only able to predict "up" days. Thus, M2 has some unreliable results in terms of predictive abilities. On a general note we find that M8 predicts 4,8% better, with 13,4% better at ”down” days and 1% better at ”up” days. Overall it established 14,4% better AUC. On a side note, M8 have weaknesses in normal market conditions compared with the best predictive abilities taken from each model. Though it is close to optimal values slacking a mere 96,4% of total accurate predictions, predicting ”down” days with 92,2% of max, with only 99,6% of maximal value of ”up” days predictions and 96,9% in terms of the best AUC observation.

4.5 General effect of different explanatory variables

Table 4.5.1 Data output of M3

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)        4.283e-01  2.094e-01   2.045  0.04081 *  
omx_return        -1.033e+01  4.749e+00  -2.175  0.02966 *  
indicator_variable -3.357e-01  1.430e-01  -2.347  0.01893 *  
omx_vol            1.340e-11  1.181e-11   1.134  0.25670    
vix               -1.434e-02  8.387e-03  -1.710  0.08729 .  
stibor            -1.177e-01  4.268e-02  -2.757  0.00583 ** 
omx_range_norm     1.327e+01  8.044e+00   1.650  0.09899 .  
Treturn           -7.233e-02  1.819e-01  -0.398  0.69083    
Vixreturn         -5.707e+00  8.379e-01  -6.811 9.71e-12 ***
StiborR           -1.863e-01  1.340e+00  -0.139  0.88945    
Rang.n.R          -9.574e-02  1.150e-01  -0.832  0.40518    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In table 4.5.1 we see which main explanatory variables are significant without considering any lags. Regarding significance, we find that the probability of “up” has a negative relationship with the percentage change of OMXS30 (return_omx), indicator variable, stibor and the percentage change of Vix (Vixreturn). If we begin with the return of omx and the indicator variable, these two have the same information content (since indicator variable = 1 if omx_return > 0 and 0 if omx_return < 0). This means that if today is “up” decreases the probability of an “up” tomorrow. Regarding stibor rates we find that lower stibor rates increase the probability of an “up” day tomorrow. Regarding the Vix return, it should be interpreted that if Vix increased from yesterday to today, this decreases the probability of an “up” tomorrow.

Regarding less significant variables (one dot significant variables 0.1> p>0.05) we find that low Vix prices and smaller price ranges in percent encourage more “up”:s tomorrow. We might as well assume effects from insignificant variables, though we cannot be sure about the effects due to the poor fit.

4.6 Misclassification Analysis

As can be found in appendix 6.2 all the variables pass the multicollinearity test of VIF being < 5. One variable that is dangerously close to being excluded is daily price range in percent (*omx_range_norm)* yielding a VIF of 4.985244 in M1. This variable is not used in M8, though having one dot significance in M3 and being insignificant in M1. Thus, the effect from this variable is small, not noted in analysis and can be considered of not damaging the overall variable choice. In M3, where one dot significance was reached, the variable reached a VIF of 3.988153, indicating that the overall high VIF value in M1 is not posing any problems.

Regarding the Durbin Watson test it has been conducted on model M1, since model M1 contained all the variables. Thus, if errors are correlated in M2-M8, this would be discovered in M1. The D-W statistic of M1 is 2.043679 yielding a p-value of 0.426. We find an autocorrelation of -0.02245721 and can conclude that errors are not correlated.

Chapter five: “Predicting Tomorrow’s Direction of the Swedish Stock Market”

5. Summary, Suggestions for Further Research, Conclusion

5.1 Summary

As we conclude our model M8 improves compared to M1 and compared to the mean of all other models. What M8 basically tells us is that the response of the stock index OMXS30 being up tomorrow is highly dependent of two variables: if today is a “down” day and if the return of Vix is negative. Guessing the other direction would imply inverting these conditions. During the 8 years observed M8 achieved average correct predictions of 56,1% tested in 3 different market conditions. In normal market conditions this increases to 58,5% and in stressed conditions (R1&R2) this decreases on average to 54,9%.

In general, we find AUC values with poor discrimination. This might be due to different reasons. Either the model selections are poor or financial data is hard to predict (due to market noise). We would have wished for average AUC:s > 0.7 and higher predictive correctness. Not achieving this might be due to variable choice and/or more variables which bear significance may be needed.

As mentioned earlier M8 lacks descriptive qualities regarding all explanatory variables and what impact they have on the response. Yet, M8 contains the most important variables and outperforms the other models in different tests. It does pass the comparison with the best values of M1 and M3-M7 in general just not in normal market conditions though being close to optimal. Furthermore, it should be regarded that M8 should endure future unknown market conditions, due to the fact it performed well in R1&R2. Though, seasonality of markets can change and make this model useless. Yet, the model has good preconditions to withstand turbulent conditions since these market swings in R1&R2 are the most volatile market swings in the last 15 years.

Quite interestingly, the VIX index do not pose significance (in M1), but the percentage changes of VIX (in M1&M8) and the first lag of VIX (in M1) qualify for 5% significance. These two indicators have negative coefficients implying that if VIX rises today or rose yesterday, this increases the probability of “down” tomorrow.

It is generally assumed that the Swedish stock index follows the American one closely (see figure 3.3.3). In retrospect, VIX is a variable that measures implied expectations of price developments in the near term. Therefore, VIX is built to have predictive abilities. This index does a good job predicting outcomes, particularly “down” days.

5.2 Suggestions for Further Research

These variables have different classes. For instance, if we define VIX < 27,30$ this accounts for >85% of data during the 8 years studied. Having a common/rare or low/high parameter on variables such as VIX, STIBOR, Range, Volumes & OMXS30 can improve the model. The values of these variables have descriptive information that can be assessed by the market empirically. For instance, it’s generally regarded that low STIBOR rates and low VIX prices encourage stock market appreciation.

One could as well include more dummy variables of other factors not mentioned in this report. One could look at leading stocks as predictors of OMXS30. There could as well be another classification of the response variable, i.e. to classify the response in different regressions if the return is +0,5%, +1%, +1,5%, +2% ... and vice versa with negative returns. Then one could achieve several models that captures the predictive outcome in a more nuanced manner.

There could as well be a study on other stock indices, to compare if there is a possibility of predicting “up”:s or “down”:s in other markets. One could as well do the inverse by studying individual stocks by having inputs on OMXS30 and using index data to predict stock movements.

Since OMXS30 correlates highly with S&P500 a natural study would be to see how data from American stock markets can predict Swedish stock returns.

Studying binary indicators created from changes instead of studying actual changes, has some benefits in a statistical sense. As seen in figure 3.3.2, we find that the 30 day average of “down”:s and “up”:s, asserted as number of “0”:s and “1”:s, is distributing in a systematic mode. We can find extreme values where in a medium-term perspective should render good probabilities of finding reversals in trends. Therefore, the systematic approach of working with dummy variables might expand models built on forecasting and predicting.

5.3 Conclusion

We succeed establishing a model which persistently predicts with a correctness of > 50%. In a general sense, we can distinguish the model from guessing, yet we note that all models are below acceptable discrimination. The ROC estimate is not simply high enough making us conclude the models presented have poor discrimination.

Regarding the ability to predict stock market outcomes with statistical confidence either takes different approaches, different models or different data. The validity of predictions is as well a hot topic of discussion. A model being able to predict during a longer season, say 10 years, may fail the next decade. Regarding uncertainty, we assume that testing data will be similar to training data. This assumption is not in accordance with a dynamic market ever evolving. Yet, searching for edges in predictability of the stock market seems possible and doable. Thus, these kinds of models might prove useful for exploring the unknown. Maybe, with some improvements, predicting tomorrow’s outcome could reach acceptable discrimination.

References:

Anon., 2004. *The Oxford English Dictionary.* Draft Revision March 2004 red. u.o.:u.n.

Avellaneda, M. L. A. &. P. A., 1995. *Willmott Wiki.* [Online]
Available at: http://www.wilmottwiki.com/wiki/index.php?title=Volatility
[Used 11 October 2016].

Bourquin, T., 2007. *Using Volume To Predict Price Movement,* North Palm Avenue Sarasota: MoneyShow.

Brenner, M. & Fand Galai, D., 1989. New Financial Instruments for Hedging Changes in Volatility. *Financial Analysts Journal,* Issue July/August.

CBOE, 2016. *www.cboe.com.* [Online]
Available at: http://www.cboe.com/micro/vix/historical.aspx
[Used 13 October 2016].

Everitt, B., 1998. The Cambridge Dictionary of Statistics. *Cambridge University Press,* Issue Cambridge, UK New York:.

Hosmer, D. W. & Lemeshow, S., 2000. *Applied Logistic Regression.* Second Edition red. u.o.:John Wiley & Sons, Inc..

Investopedia, 2016. *Investopedia.* [Online]
Available at: http://www.investopedia.com/terms/v/volume.asp
[Used 11 October 2016].

Investopedia, 2016. *Range.* [Online]
Available at: http://www.investopedia.com/terms/r/range.asp
[Used 11 October 2016].

Investopedia, 2016. *www.investopedia.com.* [Online]
Available at: http://www.investopedia.com/articles/economics/09/lehman-brothers-collapse.asp
[Used 14 October 2016].

Longford, N. T., 1986. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *BIOMETRIKA,* 74(Fourth Edition), pp. 817-827.

MathWorks, 2016. *https://se.mathworks.com.* [Online]
Available at: https://se.mathworks.com/help/econ/examples/time-series-regression-viii-lagged-variables-and-estimator-bias.html?requestedDomain=www.mathworks.com
[Used 12 October 2016].

Mood, A. & Graybill, F., 1963. *Introduction to the Theory of Statistics.* 2nd edition red. u.o.:McGraw-Hill.

NasdaqOMXNordic, 2016. *www.nasdaqomx.com.* [Online]
Available at:

http://www.nasdaqomx.com/transactions/trading/fixedincome/fixedincome/sweden/stiborswaptreasuryfixing
[Used 11 October 2016].

NasdaqOMXNordic, 2016. *www.nasdaqomxnordic.com.* [Online]
Available at:
http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0000337842
[Used 13 October 2016].

Rao, C. R., 1973. *Linear Statistical Inference and Its Applications.* Second Edition red. New Yor: Wiley, Inc..

Riksbanken, 2016. *www.riksbanken.se.* [Online]
Available at: http://www.riksbank.se/sv/Rantor-och-valutakurser/Sok-rantor-och-valutakurser/
[Used 13 October 2016].

San Diego State University, 2013. *Logistic Regression (R),* San Diego: Wordpress.

Sveriges Riksbank, 2012. *The Riksbank's review of Stibor,* Stockholm: Sveriges Riksbank.

Tague, N. R., 2004. *www.asq.org.* [Online]
Available at: http://asq.org/learn-about-quality/seven-basic-quality-tools/overview/overview.html
[Used 17 October 2016].

The Minitab Blog, 2016. *What are the effects of Multicollinearity and When can I ignore them?,* Pennsylvania: Minitab Inc.

www.mathworks.com, 2016. *www.mathworks.com.* [Online]
Available at: www.mathworks.com
[Used 11 Augusti 2016].

6. Appendices:

6.1 Tableplot and output data for M1-M8

Table 6.1.1 Tableplot of the Dataset

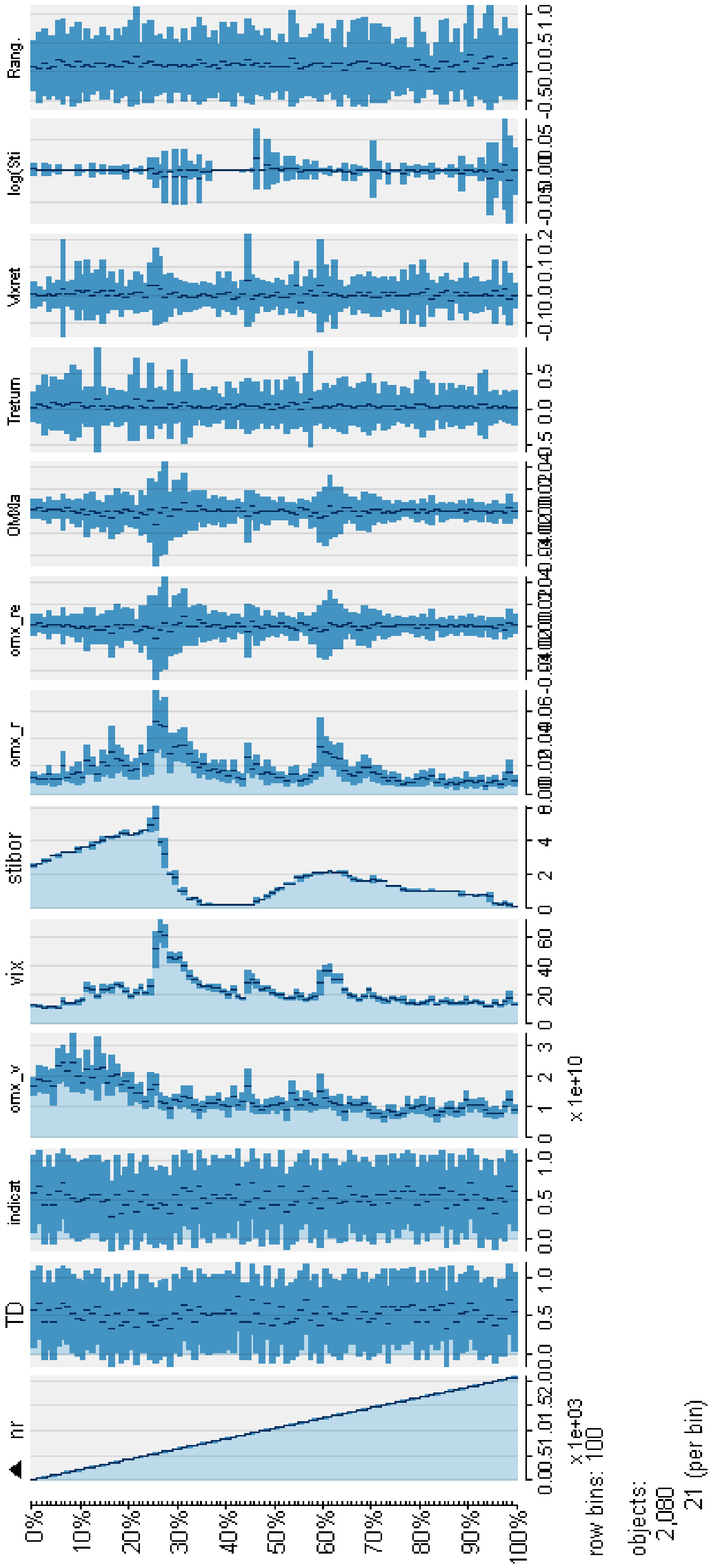


Table 6.1.2 Output of M1-M8

	ACC.P%	PDOWN	PUP	AUC
M1	0.606715	0.61111	0.604396	0.63263
M1R1	0.47482	0.42384	0.50376	0.46438
M1R2	0.48201	0.42754	0.50896	0.459
MEANM1	0.52118	0.4875	0.53904	0.51867
	ACC.P%	PDOWN	PUP	AUC
M2	0.52632	NAN	0.52632	0.58384
M2R1	0.51675	0.493421	0.578947	0.56217
M2R2	0.54545	0.833333	0.536946	0.52461
MEANM2	0.52951	0.66338	0.5474	0.55687
	ACC.P%	PDOWN	PUP	AUC
M3	0.58513	0.59504	0.58108	0.62042
M3R1	0.506	0.46795	0.52874	0.48742
M3R2	0.4964	0.43966	0.51827	0.47551
MEANM3	0.52918	0.50088	0.5427	0.52778
	ACC.P%	PDOWN	PUP	AUC
M4	0.59472	0.59854	0.59286	0.62815
M4R1	0.4988	0.45638	0.52239	0.49243
M4R2	0.46763	0.39844	0.49827	0.45713
MEANM3	0.52038	0.48445	0.53784	0.5259
	ACC.P%	PDOWN	PUP	AUC
M5	0.57554	0.57037	0.57801	0.61312
M5R1	0.51079	0.47297	0.5316	0.49134
M5R2	0.47482	0.40336	0.50336	0.47354
MEANM5	0.52038	0.48223	0.53766	0.526
	ACC.P%	PDOWN	PUP	AUC
M6	0.58034	0.5814	0.57986	0.61836
M6R1	0.50839	0.4698	0.52985	0.48853
M6R2	0.47722	0.40833	0.50505	0.4765
MEANM6	0.52198	0.48651	0.53825	0.5278
	ACC.P%	PDOWN	PUP	AUC
M7	0.58034	0.55932	0.59583	0.61945
M7R1	0.4988	0.4586	0.52308	0.48871
M7R2	0.49161	0.45294	0.52998	0.47756
MEANM7	0.52358	0.49029	0.54963	0.52857
	ACC.P%	PDOWN	PUP	AUC
M8	0.58513	0.56354	0.60169	0.61296
M8R1	0.57314	0.65672	0.53357	0.6275
M8R2	0.52518	0.49419	0.54694	0.56762
MEANM8	0.56115	0.57148	0.56073	0.60269

Table 6.1.3 Prediction output of M1-M8

			PRED	
	PRED DOWN/DOWN	PRED DOWN/UP	UP/UP	PRED UP/DOWN
M1	88	56	165	108
M1R1	64	87	134	132
M1R2	59	79	142	137

			PRED	
	PRED DOWN/DOWN	PRED DOWN/UP	UP/UP	PRED UP/DOWN
M2	NAN	NAN	110	99
M2R1	75	77	33	24
M2R2	5	1	109	94

			PRED	
	PRED DOWN/DOWN	PRED DOWN/UP	UP/UP	PRED UP/DOWN
M3	72	49	172	124
M3R1	73	83	138	123
M3R2	51	65	156	145

			PRED	
	PRED DOWN/DOWN	PRED DOWN/UP	UP/UP	PRED UP/DOWN
M4	82	55	166	114
M4R1	68	81	140	128
M4R2	51	77	144	145

			PRED	
	PRED DOWN/DOWN	PRED DOWN/UP	UP/UP	PRED UP/DOWN
M5	77	58	163	119
M5R1	70	78	143	126
M5R2	48	71	150	148

			PRED	
	PRED DOWN/DOWN	PRED DOWN/UP	UP/UP	PRED UP/DOWN
M6	75	54	167	121
M6R1	70	79	142	126
M6R2	49	71	150	147

			PRED	
	PRED DOWN/DOWN	PRED DOWN/UP	UP/UP	PRED UP/DOWN
M7	99	78	143	97
M7R1	72	85	136	124
M7R2	77	93	128	119

			PRED	
	PRED DOWN/DOWN	PRED DOWN/UP	UP/UP	PRED UP/DOWN
M8	102	79	142	94
M8R1	88	46	151	132
M8R2	85	87	134	111

6.2 VIF Table & CrPlots

Table 6.2.1 VIF Table M1

```
> vif(model_test)

   omx_return indicator_variable   omx_vol

    2.527460        2.059900        2.105991

      vix          stibor    omx_range_norm

    3.708813        1.558812        4.985244

    OMXlag1         OMXlag2         OMXlag3

    1.524654        1.530164        1.533449

    OMXlag4         OMXlag5         Treturn

    1.515086        1.399877        1.801263

      Tlag1         Tlag2         Tlag3

    1.865118        1.811436        1.783480

      Tlag4         Tlag5      Vixreturn

    1.720346        1.470915        1.484723

    Vixlag1         Vixlag2         Vixlag3

    1.651683        1.671957        1.653225

    Vixlag4         Vixlag5         StiborR

    1.629964        1.521415        1.028722

    StiborL1        StiborL2        StiborL3

    1.026374        1.023463        1.025095

    StiborL4        StiborL5        Rang.n.R

    1.026951        1.022189        2.817675

    Rang.n.L1       Rang.n.L2       Rang.n.L3

    2.490097        2.333397        2.248221

    Rang.n.L4       Rang.n.L5

    2.127607        1.634985
```

Figure 6.2.2 CrPlot 1

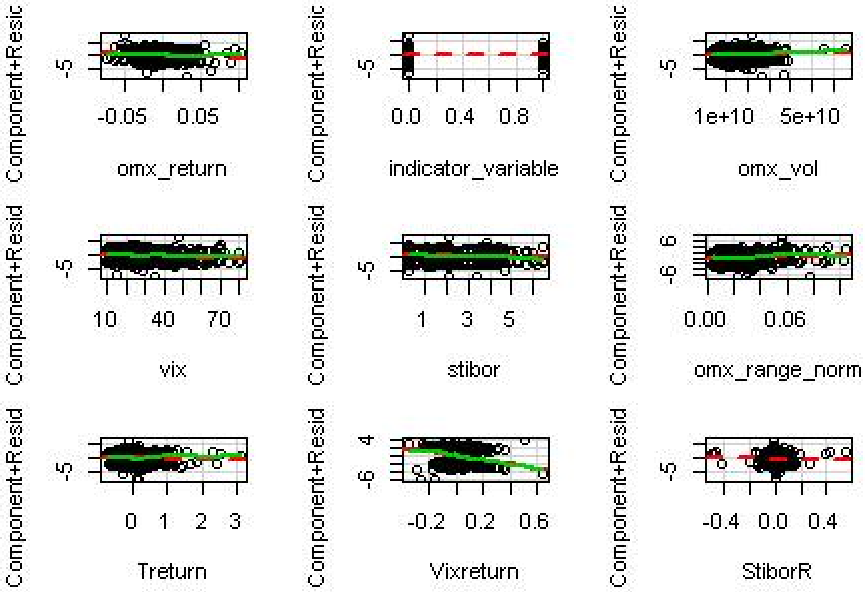


Figure 6.2.3 CrPlot 2

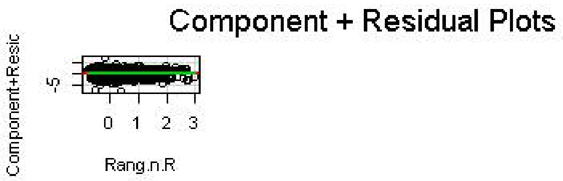


Table 6.2.4 VIF Table M3

```
omx_return indicator_variable        omx_vol

     2.383744        2.028964        1.838686

        vix          stibor    omx_range_norm

     3.240295        1.477338        3.988153

      Treturn       Vixreturn        StiborR

     1.387652        1.345732        1.013065

     Rang.n.R

     1.798848
```

6.3 Distributions and histograms

Figure 6.3.1 Range per price of OMXS30

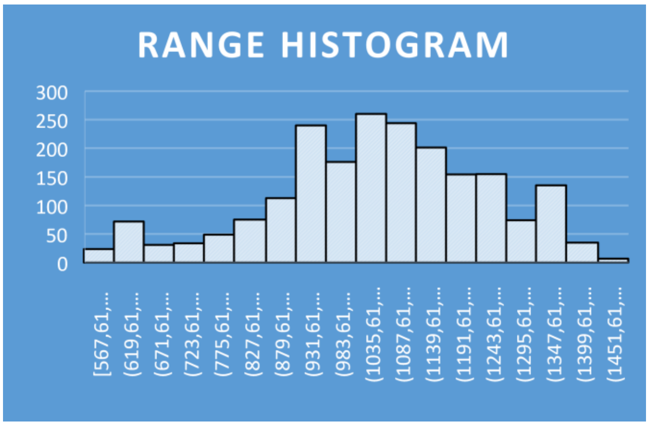


Figure 6.3.2 Range plot

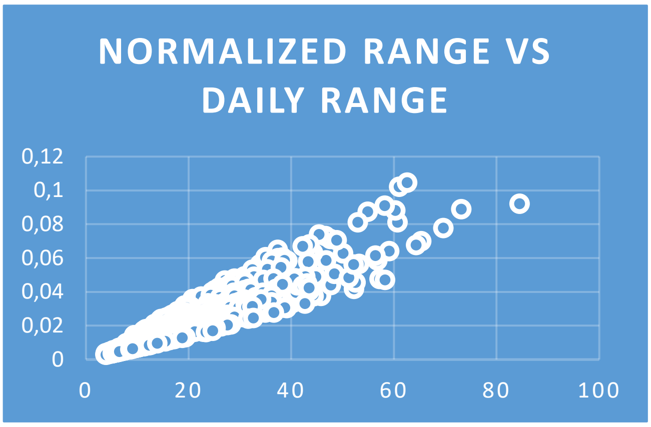


Figure 6.3.3 Stibor

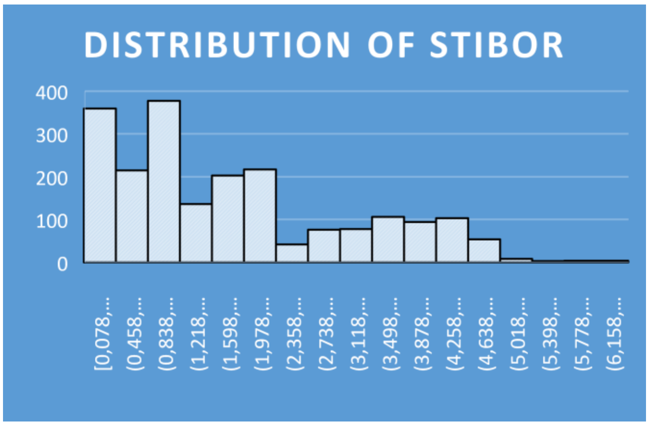


Figure 6.3.4 OMXS30 Close

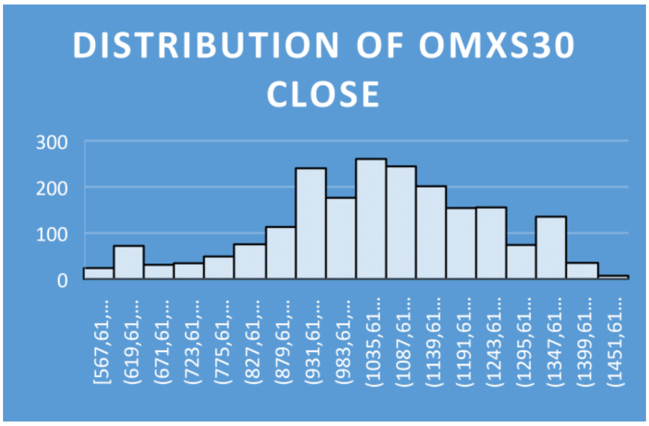


Figure 6.3.5 Volume

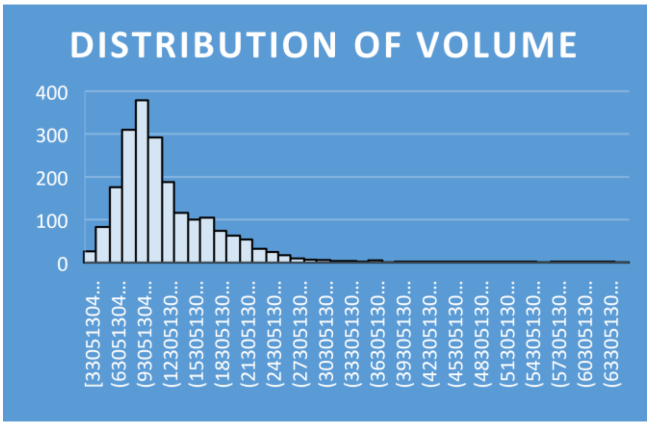


Figure 6.3.6 Vix

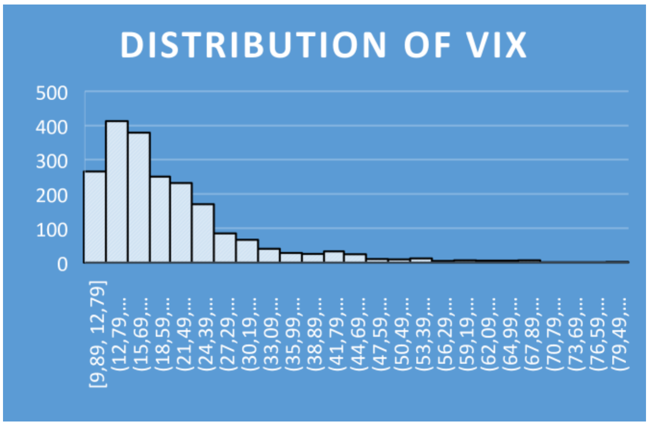


Figure 6.3.7 OMXS30 Lags

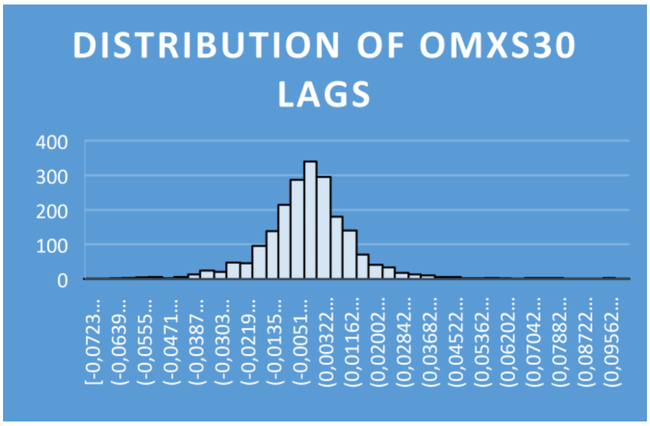


Figure 6.3.8 Volume Lags

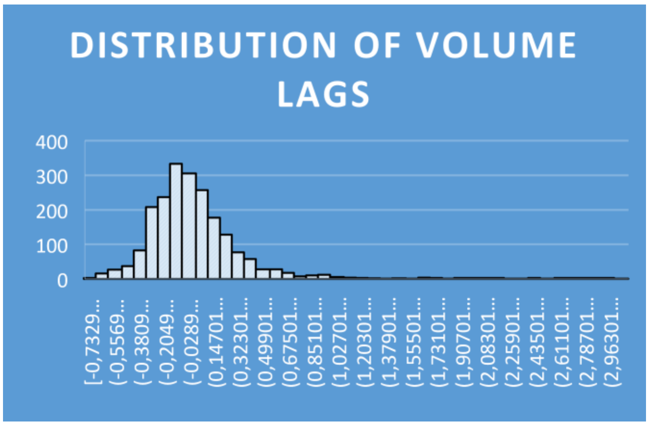


Figure 6.3.9 Vix Lags

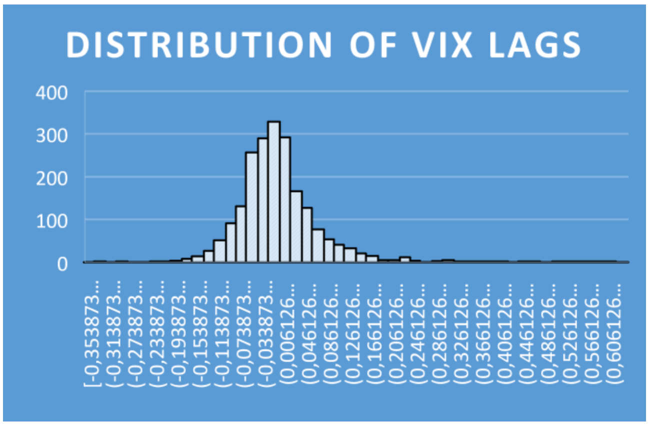


Figure 6.3.10 Stibor Lags

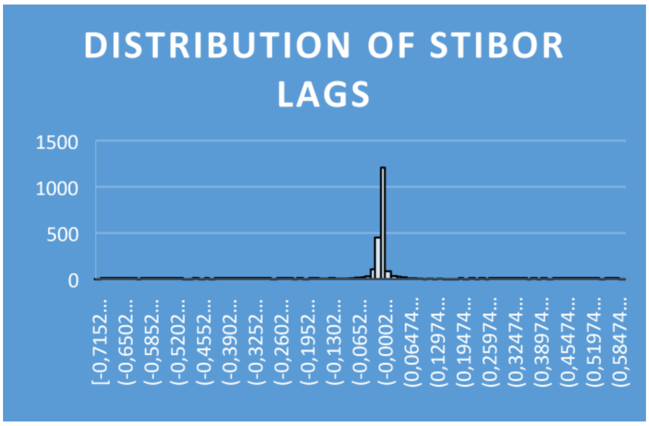
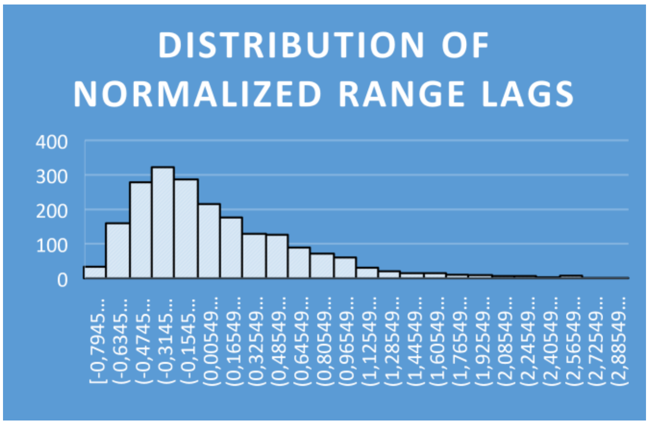


Figure 6.3.11 Normalized Range Lags

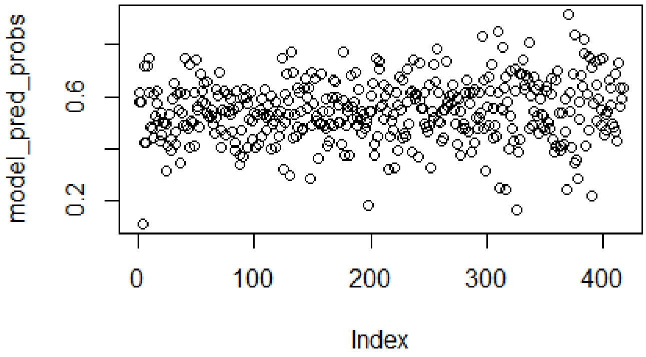


6.4 Logistic Regression outputs

Data M1:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0281  -1.1505   0.7832   1.0955   2.0107

Coefficients:
                    Estimate   Std.Error   zvalue Pr(>|z|)
(Intercept)        4.933e-01  2.308e-01   2.138  0.03255 *
omx_return        -1.257e+01  4.945e+00  -2.543  0.01100 *
indicator_variable -3.639e-01  1.455e-01  -2.502  0.01236 *
omx_vol            2.125e-11  1.275e-11   1.666  0.09572 .
vix               -1.497e-02  9.121e-03  -1.641  0.10075
stibor            -1.393e-01  4.432e-02  -3.143  0.00167 **
omx_range_norm     1.256e+01  9.119e+00   1.378  0.16834
OMXlag1           -1.065e+01  3.848e+00  -2.769  0.00563 **
OMXlag2           -5.248e+00  3.863e+00  -1.359  0.17425
OMXlag3           -4.382e+00  3.832e+00  -1.144  0.25274
OMXlag4           -5.693e+00  3.834e+00  -1.485  0.13758
OMXlag5            1.180e+00  3.676e+00   0.321  0.74815
Treturn           -1.140e-01  2.099e-01  -0.543  0.58713
Tlag1             -1.527e-01  2.118e-01  -0.721  0.47090
Tlag2             -8.709e-02  2.082e-01  -0.418  0.67570
Tlag3             -4.078e-01  2.093e-01  -1.949  0.05131 .
Tlag4             -2.048e-01  2.016e-01  -1.016  0.30982
Tlag5             -4.057e-01  1.881e-01  -2.157  0.03100 *
Vixreturn         -6.153e+00  8.898e-01  -6.915 4.69e-12 ***
Vixlag1           -1.811e+00  8.863e-01  -2.043  0.04101 *
Vixlag2           -5.697e-01  8.888e-01  -0.641  0.52158
Vixlag3           -1.352e+00  8.873e-01  -1.523  0.12763
Vixlag4           -3.865e-01  8.745e-01  -0.442  0.65850
Vixlag5            1.794e-01  8.510e-01   0.211  0.83306
StiborR           -3.150e-01  1.372e+00  -0.230  0.81839
StiborL1          -1.696e+00  1.376e+00  -1.233  0.21762
StiborL2           1.911e+00  1.462e+00   1.307  0.19117
StiborL3           4.940e-01  1.341e+00   0.368  0.71264
StiborL4          -3.679e-01  1.365e+00  -0.269  0.78758
StiborL5          -1.277e+00  1.467e+00  -0.870  0.38414
Rang.n.R          -1.406e-01  1.453e-01  -0.968  0.33297
Rang.n.L1         -7.158e-02  1.357e-01  -0.527  0.59793
Rang.n.L2         -2.183e-01  1.320e-01  -1.654  0.09811 .
Rang.n.L3          8.742e-02  1.297e-01   0.674  0.50018
Rang.n.L4          7.216e-02  1.257e-01   0.574  0.56603
Rang.n.L5          1.226e-01  1.106e-01   1.109  0.26760
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2301.9  on 1661  degrees of freedom
Residual deviance: 2200.0  on 1626  degrees of freedom
AIC: 2272

Number of Fisher Scoring iterations: 4
```

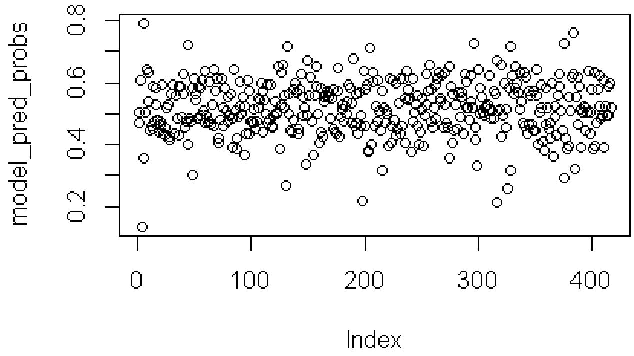
Figure 6.4.1 Logistic Regression Output M1



Data M8:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7822  -1.1810   0.8817   1.1253   2.1563

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.35568    0.07529   4.724 2.31e-06 ***
indicator_variable -0.52054    0.10648  -4.889 1.01e-06 ***
vixreturn          -5.15515    0.77091  -6.687 2.28e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2301.9  on 1661  degrees of freedom
Residual deviance: 2245.2  on 1659  degrees of freedom
AIC: 2251.2

Number of Fisher Scoring iterations: 4
```

Figure 4.3.2 Logistic Regression Output M8



6.5 R Code

The following code has been reduced to size 8 to save number of pages. This code is for the first model only with robustness tests, remaining models are done with the same procedure but with different equations in the GLM command on line 18,63 & 108.

```
"full model"
"START"
attach(rdata5)
library(rcompanion)
library(lordif)
library(ROCR)
"summary(rdata5)"
"cor(rdata5[,-1 -2 -12 -13])"
T=1663
training = (nr < T)
testing = !training
training_data = rdata5[training,]
testing_data = rdata5[testing,]
direction_testing=indicator_variable[training]
output = factor(TD)
output_test = TD[testing]
plot(output_test)
model_test=glm(TD~omx_return+indicator_variable+omx_vol+vix+stibor+omx_range_norm+OMXlag1+OMXlag2+OMXlag3+OMXlag4+OMXlag5+Treturn+Tlag1+Tlag2+Tlag3+Tlag4+Tlag5+Vixreturn+Vixlag1+Vixlag2+Vixlag3+Vixlag4+Vixlag5+StiborR+StiborL1+StiborL2+StiborL3+StiborL4+StiborL5+Rang.n.R+Rang.n.L1+Rang.n.L2+Rang.n.L3+Rang.n.L4+Rang.n.L5, data = training_data, binomial, control = list(maxit=25))
summary(model_test)
model_pred_probs=predict(model_test,testing_data,type = "response")
summary(model_pred_probs)
plot(model_pred_probs)
model_pred_TD=rep("0", 2080-T)
model_pred_TD[model_pred_probs > 0.5] = "1"
plot(model_pred_TD)
table(model_pred_TD,output_test)
1-mean(model_pred_TD!=output_test)
A=table(model_pred_TD,output_test)
A[1]/(A[1]+A[3])
1-A[2]/(A[2]+A[4])
mean(output_test)
sd(model_pred_probs)
"goldmann statistic"
sd(model_pred_probs)*(1-mean(model_pred_TD!=output_test))
pr <- prediction(model_pred_probs,output_test)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
```

```
auc
nagelkerke(model_test)
"require(tabplot)
require(ggplot2)
tableplot(rdata5)"

"full model" "ROBUST1"
"START"
attach(rdata5)
library(rcompanion)
library(lordif)
library(ROCR)
"summary(rdata5)"
"cor(rdata5[,-1 -2 -12 -13])"
T=1663
training = (NR < T)
testing = !training
training_data = rdata6[training,]
testing_data = rdata6[testing,]
direction_testing=indicator_variable[training]
output = factor(TD)
output_test = TD[testing]
plot(output_test)
model_test=glm(TD~omx_return+indicator_variable+omx_vol+vix+stibor+omx_range_norm+OMXlag1+OMXlag2+OMXlag
3+OMXlag4+OMXlag5+Treturn+Tlag1+Tlag2+Tlag3+Tlag4+Tlag5+Vixreturn+Vixlag1+Vixlag2+Vixlag3+Vixlag4+Vixlag5+Stib
orR+StiborL1+StiborL2+StiborL3+StiborL4+StiborL5+Rang.n.R+Rang.n.L1+Rang.n.L2+Rang.n.L3+Rang.n.L4+Rang.n.L5,
data = training_data, binomial, control = list(maxit=25))
summary(model_test)
model_pred_probs=predict(model_test,testing_data,type = "response")
summary(model_pred_probs)
plot(model_pred_probs)
model_pred_TD=rep("0", 2080-T)
model_pred_TD[model_pred_probs > 0.5] = "1"
plot(model_pred_TD)
table(model_pred_TD,output_test)
1-mean(model_pred_TD!=output_test)
A=table(model_pred_TD,output_test)
A[1]/(A[1]+A[3])
1-A[2]/(A[2]+A[4])
mean(output_test)
sd(model_pred_probs)
"goldmann statistic"
sd(model_pred_probs)*(1-mean(model_pred_TD!=output_test))
pr <- prediction(model_pred_probs,output_test)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
```

```
plot(prf)
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
nagelkerke(model_test)
"require(tabplot)
require(ggplot2)
tableplot(rdata5)"

"full model" "Robust2"
"START"
attach(rdata5)
library(rcompanion)
library(lordif)
library(ROCR)
"summary(rdata5)"
"cor(rdata5[,-1 -2 -12 -13])"
T=1663
training = (NR < T)
testing = !training
training_data = rdata7[training,]
testing_data = rdata7[testing,]
direction_testing=indicator_variable[training]
output = factor(TD)
output_test = TD[testing]
plot(output_test)
model_test=glm(TD~omx_return+indicator_variable+omx_vol+vix+stibor+omx_range_norm+OMXlag1+OMXlag2+OMXlag
3+OMXlag4+OMXlag5+Treturn+Tlag1+Tlag2+Tlag3+Tlag4+Tlag5+Vixreturn+Vixlag1+Vixlag2+Vixlag3+Vixlag4+Vixlag5+Stib
orR+StiborL1+StiborL2+StiborL3+StiborL4+StiborL5+Rang.n.R+Rang.n.L1+Rang.n.L2+Rang.n.L3+Rang.n.L4+Rang.n.L5,
data = training_data, binomial, control = list(maxit=25))
summary(model_test)
model_pred_probs=predict(model_test,testing_data,type = "response")
summary(model_pred_probs)
plot(model_pred_probs)
model_pred_TD=rep("0", 2080-T)
model_pred_TD[model_pred_probs > 0.5] = "1"
plot(model_pred_TD)
table(model_pred_TD,output_test)
1-mean(model_pred_TD!=output_test)
A=table(model_pred_TD,output_test)
A[1]/(A[1]+A[3])
1-A[2]/(A[2]+A[4])
mean(output_test)
sd(model_pred_probs)
"goldmann statistic"
```

```
sd(model_pred_probs)*(1-mean(model_pred_TD!=output_test))
pr <- prediction(model_pred_probs,output_test)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
nagelkerke(model_test)
"require(tabplot)
require(ggplot2)
tableplot(rdata5)"
red_model_test=glm(TD~omx_return+indicator_variable+omx_vol+vix+stibor+omx_range_norm+OMXlag1+OMXlag2+OMXlag3+OMXlag4+OMXlag5+Treturn+Tlag1+Tlag2+Tlag3+Tlag4+Tlag5+Vixreturn+Vixlag1+Vixlag2+Vixlag3+Vixlag4+Vixlag5+StiborR+StiborL1+StiborL2+StiborL3+StiborL4+StiborL5+Rang.n.R+Rang.n.L1+Rang.n.L2+Rang.n.L3+Rang.n.L4+Rang.n.L5, data = training_data, binomial, control = list(maxit=25))
backwards= step(model_test)
summary(backwards)
"dropping omxlag4,stiborlag2,rangelag2,tlag5,tlag3"
back2= glm(TD~omx_return+indicator_variable+omx_vol+stibor+OMXlag1+Vixreturn+Vixlag1, data = training_data, binomial, control = list(maxit=25))
summary(back2)
```