Trading volume at Avanza

GRETA KNUTSSON
KAMYAR ESPAHBODI
Trading volume at Avanza

GRETA KNUTSSON

KAMYAR ESPAHBODI
Abstract

Producing a model explaining the trading volume can be attractive for companies who’s main revenue resides on it. Previous studies have shown that factors such as stock returns, volatility and uncertainty affects the trading volume.

The purpose of this work is to clarify the consensus that prevails and determine the factors that impact Avanza’s customers trading volume. Factors such as daily stock returns and economic, political and financial uncertainty are analyzed through a multiple linear regression analysis with a daily time period between 2000-2019. The work is thus designed within the framework of mathematical statistics and industrial economics.

To be able to draw a conclusion, further investigation is required in the form of a time series analysis in combination with a deeper understanding of the applied area and the mathematical methods that have been used.

Keywords

Abstract

Att ta fram en modell som förklarar handelsvolymen kan vara eftertraktat hos företag vars huvudintäkter beror av den. Tidigare forskning visar att faktorer som prisförändringar på aktiemarknaden, volatilitet och osäkerhet påverkar handelsvolymen.


För att kunna dra en slutsats krävs vidare undersökning i form av en tidsserieanalys och en djupare förståelse av det tillämpade området och metoderna som har använts.

Nyckelord

Kandidatexamensarbete, Regressionsanalys, Handelsvolym, Politisk Osäkerhet, Börsindex, Avanza
Acknowledgements

We would like to thank our supervisor at the Royal Institute of Technology, Camilla Johansson Landén from the Department of Mathematics, and Robert Ingemarsson and Rasmus Åkerblom from Avanza Holding Bank AB, whom assisted us with relevant data and supported us throughout this thesis.
Authors
Kamyar Espahbodi <kamyare@kth.se> and Greta Knutsson <gretak@kth.se>
Industrial Engineering and Management
KTH Royal Institute of Technology
Applied Mathematics TMAI

Place for Project
Stockholm, Sweden

Supervisors
Camilla Johansson Landén
Lindstedtsvägen 25, 114 28, Stockholm
KTH Royal Institute of Technology

Julia Liljegren
Lindstedtsvägen 30, 114 28, Stockholm
KTH Royal Institute of Technology

Robert Ingemarsson
Regeringsgatan 103, 111 39, Stockholm
Avanza Bank Holding AB
# Contents

1 Introduction ........................................ 1
   1.1 Background ........................................ 1
   1.2 Problem ............................................ 2
   1.3 Economic Theory .................................... 2
   1.4 Goal and Purpose ................................... 4
   1.5 Methodology ....................................... 6
   1.6 Stakeholder Analysis ............................. 6
   1.7 Scope ................................................ 7

2 Regression Analysis ............................... 9
   2.1 Multiple Regression ............................... 9
   2.2 Ordinary Least Squares (OLS) .................... 10
   2.3 OLS Assumptions ................................... 11
   2.4 Residual analysis .................................. 14
   2.5 Leverage and Influence ......................... 17
   2.6 Variable Selection ................................ 19
   2.7 Transformations ................................... 25

3 Data .................................................. 27
   3.1 Trading Volume .................................... 27
   3.2 Equity .............................................. 28
   3.3 Volatility .......................................... 29
   3.4 Economic Policy Uncertainty (EPU) ............ 29

4 Result ............................................... 31
   4.1 Residual Analysis ................................. 31
   4.2 Leverage and influential points ................. 32
   4.3 Transformations ................................... 35
   4.4 Variable Selection ................................ 36
   4.5 Final Model ........................................ 44

5 Discussion .......................................... 46
   5.1 Methodology ....................................... 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2 Model Adequacy</td>
<td>48</td>
</tr>
<tr>
<td>5.3 Previous studies</td>
<td>48</td>
</tr>
<tr>
<td>5.4 Future Work</td>
<td>49</td>
</tr>
<tr>
<td>5.5 Avanza</td>
<td>49</td>
</tr>
</tbody>
</table>

6 Conclusion 51

References 52
1 Introduction

A traditional stockbroker is a person who executes buy and sell orders on the behalf of their client. Nowadays, the influence of internet has enabled the stockbroker to carry out orders at a lower commission rate than before, hence the birth of the discount broker. The discount broker’s business is as such online based, which implies low overhead costs and thus reduced commission rates and fees for the client. The chosen strategy for battling the market competition being high volume and low cost, the broker’s main income, in a broad sense, mainly depends on the number of client transactions. Therefore, it lies in the broker’s interest to fully understand the behaviour of the driving factors behind their customers trading volume, since it directly correlates with their aggregated commission revenue.

1.1 Background

Avanza Bank Holding AB, formally founded 1999, is Sweden’s largest online stock broker with over 800,000 customers. Its business idea is to, through low charges, offer a broad variety of saving products and educational and supporting services within the area of personal saving in Sweden. They also offer market competitive mortgage loans and pension solutions and have been ranked as having the happiest customers nine times in a row for each year within the industry.

As any other online stock broker, Avanza takes a share of each customers trade, a commission, which constitutes their primary source of income. Hence, their aggregated revenue is directly dependent on their trading volume, i.e. the number of shares transacted every day. By understanding the underlying mechanisms affecting their customers trading volume, it is possible for Avanza optimize their resources through forecasting or developing new products based on the special behaviour, streamlining their organization.

---

1 Avanza®, History
2 Avanza Bank Holding AB, Årsredovisning 2018
3 Svenskt Kvalitetsindex, Personlig service utmanar digitala tjänster
1.2 Problem

However, Avanza’s customers trading behaviour does not solely depend on internal factors, such as new user interface improvements or new functionality features, in fact, their financial result is affected by market cyclical effects such as stock market developments, volatility and federal funds rate. Avanza states, in their annual report, that the previous year 2018 was characterized by political uncertainty and that SIX return index (an index representing all of the stocks on the Swedish market) decreased by 8% and that the number of transactions rose by 15% and the revenue by 8%, compared to the previous year [2017].

It therefore lies in Avanza’s interest to know how much power they have in controlling their own revenue stream, i.e. is there still room for internal innovations and new product features that could potentially further raise the commission revenue, or does it, to some extent, only depend on exogenous factors? In other words, how much can they affect their revenue stream, and to what extent?

1.3 Economic Theory

Although no previous studies within the area was found, some key insights were gained by studying previous literature. The trading activity of a population is a complex phenomenon comprised of several factors both physiological, such as investor overconfidence, and systematic, such as calendar effects. The following sub sections further investigates some of these factors.

1.3.1 Trading Volume

It is hard to predict future trading volumes, i.e. the number of shares being transacted at a certain time point, because of its complexity and dependency on human behaviour. In general, the trading volume is driven by at least two forces; changes in heterogeneity of beliefs and the disposition effect. Summarized, these empir-

---

4 Avanza Bank Holding AB, Årsredovisning 2018, pp. 49-50
5 Shefrin, A Behavioral Approach to Asset Pricing
ically supported theories states that the investors overconfidence about their valuation and trading skills can explain the high observed trading volume.\textsuperscript{6}

Since these drivers are relatively subjective and hard to measure, from a quantitative approach, the psychological and individual factors of human being, such as overconfidence, personal beliefs, irrationality, are disregarded in this study.

\subsection*{1.3.2 Factors}

Focusing on more numerically obtainable factors, several stand out.

Several studies have shown that there exists a positive correlation between the price changes of stocks and trading volume in financial markets.\textsuperscript{7, 8, 9} A study examining potential factors affecting trading volume in European markets found the asymmetric price-volume relation in over 70 percent of the analyzed stocks.\textsuperscript{10} Based on these studies, one could assume that the price is of great importance when analyzing trading volume. Another study that analyzed the daily trading volume of the Swedish stock market found that the marketplace of the stock, contrary to belief, was irrelevant for explaining the volume, while the shareholder structure, free float and the number of outstanding shares in a company were relevant.\textsuperscript{11}

Perfect information availability is a key aspect of a perfect competition, therefore it is reasonable to assume that it affects the trading volume as well, since, for instance, positive information might persuade investors to buy a stock and vice versa. Economic Policy Uncertainty (EPU) is an index developed by researchers to generally estimate the uncertainty about fiscal or monetary policies, tax regulations, regime changes and uncertainty over electoral outcomes, in other words —everything that might raise the uncertainty regarding economic and policy outcomes. One study found that the EPU-index impacts the markets and that it has

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{6}Statman, Vorkink, and Thorley, “Investor Overconfidence and Trading Volume”
\item \textsuperscript{7}Ying, “Stock Market Prices and Volumes of Sales”
\item \textsuperscript{8}Cornell, “The Relationship between Volume and Price Variability in Future Markets”
\item \textsuperscript{9}Clark, “A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices”
\item \textsuperscript{10}Batrinca, Hesse, and Treleaven, “Examining drivers of trading volume in European markets”
\item \textsuperscript{11}Sevelin, “Swedish Stock market: Explaining trade volumes in single stocks”
\end{itemize}
\end{footnotesize}
a positive correlation with stock price volatility and reduced investment.\textsuperscript{12}

The transaction cost is another factor to consider. One study found that that trading volume is measurably responsive to changes in transaction costs.\textsuperscript{13}

Clearly, there are several factors that could help to explain the trading volume — even calender days affect it.\textsuperscript{14}

\subsection*{1.4 Goal and Purpose}

In conclusion, the change in trading volume is complex and depends on several factors. The challenge is to find an explanation of the behaviour of Avanza’s clients trading activity in order to answer the question whether the effects on the trading activity are exogenous or endogenous, i.e. can Avanza affect their customers trading volume — or does it mainly depend on macroeconomic, financial and political factors?

In summary, this thesis aims to answer the following:

1. Does uncertainty, in a financial, political and macroeconomic sense, along with market indicators such as stock prices and volatility indices, impact Avanza’s customers trading activity?
   (a) Which factors should be studied and why?

2. What are the possible relationships between the factors and what do they look like?
   (a) Can a model, given certain a significance level, explaining the fluctuations of the commission revenue be formulated?

3. How should possible relationships be implemented?
   (a) What can Avanza do with this information to increase their revenue?

\textsuperscript{12}Baker, Bloom, and Davis, “Measuring Economic Policy Uncertainty”
\textsuperscript{13}Epps, “The Demand for Brokers’ Services: The Relation between Security Trading Volume and Transaction Cost”
\textsuperscript{14}Sakalauskas and Krikščiūnienė, “The Impact of Daily Trade Volume on the day-of-the-week Effect in Emerging Stock Markets”
If the stated questions above are answered and assuming that there exists possible relationships, Avanza will have more information about their customers compared to other competitors in the same business. Therefore, the information can be seen as a tool for competitiveness.

The competitiveness for Avanza’s business can, for example, be analyzed through Michael Porter’s *Five Forces*-model, where an increase in a company’s competitiveness possibly also implicates an increasing of the revenue. The five forces in Porter’s model are:

1. Competition in the industry
2. Power of customers
3. Power of suppliers
4. Threats of substitutes
5. Potential of new entrants in the industry

By analyzing each and one of the forces, the purpose of the project can further be clarified. The first aspect of rivalry, the competition in the industry, depends on how diversified the company’s offer is, compared to other companies in the same business. The power of customer and suppliers represents customers ability to decrease prices on the output and the suppliers ability to bargain for increased prices. Threats of substitutes can be a rivalry if there exists substitutes on the market that can replace the service and goods provided by the company. Last but not least, the potential of new entrants entering the market also affects the competitiveness for the analyzed company.  

Given that the research questions are answered, Avanza will have more information on what the trading volume depends on. Therefore, they can differentiate their offer which possibly will lead to a stronger position relative to the competitors within the industry. The power of customers can decrease as the knowledge about Avanza’s customers trading behaviour increases, since the negotiation ability is usually stronger for the part with the most information.

The power of suppliers, threats of substitutes and potential new entrants in the industry can be increased by an increased knowledge on the customers’ trading behaviour. The less of an information the supplier has, the more control Avanza will have over the cost of the service and the price for the goods. The potential of new entrants entering the market can increase by an increased knowledge on the customers’ trading behaviour, since the negotiation ability is stronger for the part with the most information.

---

\(^{15}\)Chappelow, *Porter’s 5 Forces*
industry will probably remain the same, since the information mainly concerns the customers behaviour, which mainly affects the customers and the rivalry within the industry.

1.5 Methodology

By examining relevant data, such as stock price, volatility, the EPU-index etc., the purpose is to answer whether these chosen parameters are relevant in explaining the trading volume. Future research can then apply these parameters without having to investigate if they are relevant. Further, the result could help to more thoroughly define the company’s ability in manipulating its revenue stream if it mostly depends on exogenous factors.

The suggested statistical tool is multiple linear regression modeling. Extensively used within finance, regression is a statistical method to establish of the relationship between variables. In multiple regression analysis, a relationship between the response variable $y$, i.e. the variable of interest, and a set of predictor variables $\mathbf{x} = [x_1, x_2, \ldots, x_p]$ is examined.

By fitting the model with the response variable and the plausible relevant factors affecting the response variable a multiple linear model explaining the response, to some degree, can be obtained.

The literature used in the section of presenting the regression analysis is mainly *Introduction to Linear Regression Analysis* by Montgomery, Peck, and Vining.\(^{16}\)

1.6 Stakeholder Analysis

To investigate the stakeholders for this project, a stakeholder analysis is made where both internal stakeholders within the company as well as external stakeholders are included. The stakeholders can both be affected and affect the project, depending on which stage the project is in and what result comes with it.

First of all, one internal stakeholder is the management of Avanza, since it lies in

\(^{16}\)Montgomery, Peck, and Vining, *Introduction to Linear Regression Analysis*
their interest to investigate whether the trading volume depends on exogenous or endogenous behaviour. The management influences the implementation of the result, and therefore need to be informed whether any conclusions can be made. The result of this degree project can therefore be used as a tool for the management to analyze and evaluate the organization.

Second of all, the supervisors at Avanza can be seen as internal stakeholders. On one hand, the supervisors provides the project with data and knowledge which shows the importance and influence that the supervisors have on the project, while on the other hand, the project can supply the supervisors with information regarding the trading volume. This information is interesting for the supervisors at Avanza since the result can be further investigated and presented to the management of the company.

The external stakeholders does not affect the project because of the lack of awareness regarding this project. One possibly external stakeholder is the management in other companies. Even though the analyzed data is given by Avanza, there might be a general pattern representing investors all over the world and therefore it could also be useful in the management in other companies. Another possibly external stakeholder is Avanza’s customers. The customers can be affected of the project depending on what the result indicates and whether any implementations are made, but the customers will not affect the project.

1.7 Scope

Some aspects, that possibly could affect the outcome if they were examined, have been excluded in this report.

First, this study is geographically limited to Avanza’s customer that mainly reside in Sweden. Therefore, the conclusions can not be applied to trading activity behaviour in general. Second, the seasonal variation of the financial market have not been taken into consideration in this report. For example, it has been shown that there is exists a seasonal variation which affects Avanza’s trading volume (see figure 1.1).  

\[ ^{17}\text{Johanna Kull and Nicklas Andersson, Tips inför sommarbörsen} \]
Furthermore, other factors that possibly could affect the trading volume and that have been left out are the development of rents, taxes and inflation. For example, taxes and rents affect the disposable income for the households\(^\text{18}\) and therefore also the amount consumers can trade.

The inflation can affect the trading volume. If there is an increasing inflation during a specified period of time, the money will be worth less which leads to the same conclusion as for the rents and taxes, that the consumers will have less money to trade and therefore trade less than usual.

Moreover, the time frame that this thesis is delimited to is from the 3rd of January to the 29th of April 2019. This specified time period could affect the output of the project, but since Avanza was founded in year 1999,\(^\text{19}\) this aspect does possibly not have a great impact on the output.

\(^{18}\)Carlgren, *Hushållens inkomster*

\(^{19}\)Avanza®, *History*
2 Regression Analysis

Regression analysis is a statistical technique for investigating and modeling the relationship between variables.\textsuperscript{20} The word regression means "a return to a previous and less advanced or worse state, condition, or way of behaving", and may be the most widely used statistical technique.\textsuperscript{21}

The variables of interest are called response and predictor variables. The response variable is often denoted as $y$ and represents the observed outcome, given the predictor variable(s) $x$. When investigating a response variable that may be related to several regressors, a multiple regression model is appropriate.

2.1 Multiple Regression

The multiple regression model is given by the equation

$$y = X\beta + \epsilon$$

(1)

where the response $y$ consists of a vector of the response variables for each observation. The error $\epsilon$ consists of a vector of the errors for each observation. The estimation coefficients, the "betas" $\beta$, consists of a vector with beta coefficients for each predictor, where $\beta_0$ represents the intercept and $p$ the number of predictor variables. $X$ is a matrix consisting of the predictor variables.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

(2)

By estimating the betas, given the data, the fitted model produces a line along the data points through the equation

\textsuperscript{20}Montgomery, Peck, and Vining, Introduction to Linear Regression Analysis pp. 1
\textsuperscript{21}Definition of “regression” from the Cambridge Advanced Learner’s Dictionary & Thesaurus ©Cambridge University Press
\[ \hat{y} = X\hat{\beta} \]  \hspace{1cm} (3)

where \( \hat{y} \) represent the fitted values and \( \hat{\beta} \) the estimated beta coefficients.

### 2.2 Ordinary Least Squares (OLS)

The aim of the estimation of the beta coefficients is to make it as accurate as possible. The OLS-method is a popular and widely used method for obtaining the estimates \( \hat{\beta} \). OLS minimizes the sum of the squares of the differences between the observations \( y_i \) and the projected straight line and can be used to produce the estimators of the multiple regression model in Equation (1). The least-squares function is given by

\[
S(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon = (y - X\beta)^T(y - X\beta) \hspace{1cm} (4)
\]

Developing the right hand side of the equation through multiplication and taking the derivative with respect to \( \beta \) and setting it equal to zero to find the minimum, the least-squares normal equations stated in matrix notation can be written as

\[
X^T X \hat{\beta} = X^T y \hspace{1cm} (5)
\]

Or equivalently, by solving the normal equations, as

\[
\hat{\beta} = (X^T X)^{-1} X^T y \hspace{1cm} (6)
\]

Equation (6) is called the least squares estimation equation.

The Gauss-Markov Theorem states that the OLS estimator \( \hat{\beta} \) is the best linear unbiased estimator (BLUE). In other words, \( \beta \) has the smallest variance in the class of all unbiased estimators that could be produced as linear combinations of
the data.\footnote{Montgomery, Peck, and Vining, \textit{Introduction to Linear Regression Analysis}. pp. 587-588}

### 2.3 OLS Assumptions

When fulfilled, the OLS assumptions allow the OLS method to create the best possible estimates, best in the sense of having smallest variances. According to the Gauss-Markov theorem, OLS produces the best estimators when the assumptions presented down below hold. Further, the coefficient estimates $\hat{\beta}$ converge to the actual population parameters when the sample size increases to infinity and given that the assumptions are fulfilled.

#### 2.3.1 Strict Exogenity

The random errors $\epsilon$ are assumed to have mean zero, meaning that the value of one error does not depend on the value of any other regressor. This consequence follows directly from the strict exogenity assumption and can be expressed as

$$E[\epsilon|X] = 0$$

Equation (7) implies that the independent predictor $x$ is not dependent on the dependent variable $y$, i.e. an exogenous variable is able to influence the system without being influenced by it.

#### 2.3.2 Homoscedasticity

Besides having a zero mean, the errors $\epsilon$ are assumed to have the same unknown variance $\sigma^2$ in each observation, i.e. the variance is constant for each observation. The opposite is named heteroscedasticity and implies that the variance changes for different observations, which reduces the precision of the OLS generated estimators.
Investigation of the residual plots is useful for controlling the OLS assumptions and often recommended for other reasons as well, such as detecting outliers. Homoscedasticity can be investigated by plotting the residuals against the fitted values $\hat{y}_i$. By plotting the externally studentized residuals $t_i$ against the corresponding fitted value, the behaviour of the variance can be observed.

### 2.3.3 No Autocorrelation

The errors $\epsilon$ are assumed to be uncorrelated between the observations.

$$E[\epsilon_i\epsilon_j|X] = 0$$ (8)

The opposite holds if, for instance, the error term of one observation is positive and that it systematically increases the probability that the following error is positive (positive correlation). Autocorrelation is common in the context of time series data, i.e. measurements over a certain amount of time.

If autocorrelation exists the OLS estimator is still unbiased but not BLUE and the usual OLS standard errors and test statistics are no longer valid.

### 2.3.4 Normality

The normality assumptions assumes that the errors are normally distributed given the regressors, i.e.

$$\epsilon|X \sim N(\mu, \sigma^2 I_n)$$ (9)

The OLS method does not require that the normality assumption is fulfilled, however fulfillment of the assumption enables the usage of hypothesis testing, generating reliable confidence and prediction intervals.

By constructing a normal probability plot of the residuals the normality assumptions can be checked. The cumulative normal distribution is constructed such that

---

23 Montgomery, Peck, and Vining, *Introduction to Linear Regression Analysis*. pp. 136
it will plot as a straight line against the externally studentized residuals, ranked in increasing order. Since the errors are assumed to be normally distributed, the desired observation should show the points as close as possible to the straight line.

2.3.5 Multicollinearity

When multicollinearity occurs it causes lower precision of the OLS estimators. Perfect correlation occurs when one of the variables changes by a fixed proportion, the other also changes by the same fixed proportion. This indicates that the two variables are linearly dependent. Since the OLS method cannot distinguish the perfectly correlated variables a high enough correlation causes problem, i.e. multicollinearity. One could imagine multicollinearity as a two dimensional plane residing on the data points. If the data points are not orthogonal the plane will be unstable or "wiggly" and therefore vary greatly.

Since the matrix $X^TX$ consists of the dimensions $p \times p$ ($p$ being the number of predictor variables) and the $j$th column of the $X$ matrix is given by $X_j$, the matrix can be represented as $X = [X_1, X_2, ..., X_p]$. Multicollinearity is said to exist when the vectors $X_1, X_2, ..., X_p$ are linearly dependent, i.e. if there is a set of constants, $t_1, t_2, ..., t_p$ not all zero, such that

$$\sum_{j=1}^{p} t_j X_j = 0$$

(10)

Variance Inflation Factors (VIF) One way of detecting multicollinearity is to investigate the linear dependency among the regression variables, since the regressors are the columns of the $X$ matrix. By investigating the correlation matrix $C = (X^TX)^{-1}$, multicollinearity can be detected. The diagonal elements of the correlation matrix are called variance inflation factors, since they provide an index measuring the inflation of the variance of an estimated regression coefficient caused by collinearity.

24 Montgomery, Peck, and Vining, Introduction to Linear Regression Analysis. pp. 286

13
Large VIF values associated with the regression coefficients indicate that they are poorly estimated. What is meant by large values, 5 and 10 are often used based on practical experience.\textsuperscript{25} The VIF’s can be calculated by the following formula.

\[ VIF_j = C_{jj} = (1 - R^2_j)^{-1} \quad (11) \]

**Condition Number** Another way of detecting multicollinearity is by measuring the spread of the eigenvalues of $X^T X$, i.e. $\lambda_1, \lambda_2, \ldots, \lambda_p$. The condition number $\mathcal{K}$ is defined as the fraction of the largest and lowest eigenvalue.

\[ \mathcal{K} = \frac{\lambda_{\max}}{\lambda_{\min}} \quad (12) \]

$\mathcal{K} > 100$ indicates a problem with multicollinearity and $\mathcal{K} > 1000$ indicates a severe problem with multicollinearity. This is because the eigenvalues of a $p \times p$ matrix $A$, that are given by the equation $|A - \lambda I| = 0$, represent the characteristic roots of the matrix. Hence, small values of the roots indicate near-linear dependencies in the data and vice versa.

### 2.4 Residual analysis

Plotting residuals is a powerful technique for detecting outliers and estimating the overall behaviour of the model. Autocorrelation can be detected through residual plots where the plots displays residual versus time.

Residuals are defined as

\[ e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n \quad (13) \]

where $y_i$ is the observed value and $\hat{y}_i$ is the corresponding fitted value. The residual can therefore be seen as the deviation between the data and the fit. Moreover, the residual is also a measure of the variability in the response $y_i$. If any assump-\textsuperscript{25}Montgomery, Peck, and Vining, *Introduction to Linear Regression Analysis* pp. 296
tions about the residuals are wrongfully made, it will be shown in the residual analysis.

The approximate average variance of the residuals is given by

\[
\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = SS_{res} = n\frac{MS_{res}}{n-p} \tag{14}
\]

The residuals have \( n - p \) degrees of freedom, which makes the non-independence of the residuals negligible as long as the number \( n \) of residuals is not small relative to the number of coefficients \( \beta \).

### 2.4.1 Studentized Residuals

Studentized Residuals is a way of improving the residual scaling by dividing \( e_i \) by the exact standard deviation of the \( i \)th residual. The residual can be written as

\[
e = (I - H)y = (I - H)\epsilon
\]

where the hat matrix can be written as

\[
H = X(X^TX)^{-1}X^T
\]

using that \( y = X\beta + \epsilon \). Thus the covariance matrix of the residuals is

\[
\text{Var}(e) = \text{Var}(I - H)\epsilon = \sigma^2(I - H)
\]

Since the matrix \((I - H)\) generally is not diagonal, the residuals have different variances and are also correlated. The variance of the \( i \)th residual is thus \( \text{Var}(e_i) = \sigma^2(1 - h_{ii}) \), where \( h_{ii} \) is the \( i \)th diagonal element in the matrix \( H \). Now, by estimating \( \sigma^2 \) by \( MS_{Res} \), the studentized residuals are given by

\[
r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}} \tag{15}
\]
For $i$, where $i = 1, 2, ..., n$, and $\sqrt{MS_{Res}(1 - h_{ii})}$ is the standard deviation, i.e. $\sqrt{\text{Var}(e_i)}$.

Since $h_{ii}$ is the measure of the location of the $i$th point in the $x$ space, $\text{Var}(e_i)$ becomes smaller for the points $x_i$ that lie relatively far away from the center of the $x$ space. This in turn means that the studentized residuals becomes larger for these points. Thus, analyzing studentized residuals makes it easier to detect influential points which is where the violations of the basic assumptions are more likely to occur.

### 2.4.2 PRESS Residual

PRESS is generally regarded as a measure of how well a regression model will perform in predicting new data. The definition of the PRESS residual is

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}, \quad i = 1, 2, ..., n$$

(16)

Equation (16) shows that the PRESS-residual is the ordinary residual weighted with the diagonal element from the hat matrix. Taking the denominator into consideration, a large value on $h_{ii}$ implicates a small denominator which makes the deleted residual large. PRESS-residuals with small values are generally desired for a model.

### 2.4.3 R-Student

Another method to detect an outlier is the R-student. Compared to the studentized residual, R-student uses an external scaling instead of the internal which is used in the estimate of the mean square residual.

$$S_i^2 = \frac{(n - p)MS_{Res} - e_i^2/(1 - h_{ii})}{n - p - 1}$$

(17)

If an observation is influential, the estimation of the variance $S^2$ will differ significant from $MS_{Res}$. This will make the denominator close to zero and therefore
increase the residual. Now the R-student residual is given by

$$t_i = \frac{e_i}{\sqrt{S_i^2(1 - h_{ii})}}$$  \hspace{1cm} (18)

### 2.4.4 Partial Regression and Partial Residual Plots

An effective way to check the adequacy of the fit of the regression model is to make a residual plot. A partial regression plot is used in order to study the marginal relationship between the regressor given the other variables in the model. The plot shows if the assumption about the specified relationship between the response and the regressor variables has been made correctly and can also be useful for providing information about a variable that is not currently in the model. Comparing each of the regressor variables with the response, the partial regressor plot shows the marginal relation for each of the regressors.

### 2.5 Leverage and Influence

Occasionally, certain data points that negatively influences the model appear. These points are defined as influential points. The leverage point, point A in the left plot in Figure 2.1, depicts an usual $x$ coordinate, but a $y$ coordinate that seems to be located on the regression line. The $A$-point in the right plot displays an influential point with both an unusual $x$ and $y$ coordinate. As the name tells, the influential point influences the regression line to move away from the rest of data, i.e. from its more natural trajectory.

![Figure 2.1: Left: leverage point. Right: influential observation](image)

---

17
A small subset of data can negatively influence the model to such a degree that the estimates of the beta coefficients mostly rely on the subset data instead of the majority mass of the data. Since the regression model should be a construct of all of the data, these situations must be avoided. By finding and analyzing these points their impact can be determined and understood relative to the end regression model, and even removed if they are "bad".

2.5.1 Cook’s Distance

The influence of one point can be measured with respect to its location in the \( x \) space and its response. By removing one point from the sample Cook’s distance can measure its influence and this can be used as a deletion diagnostic.

\[
D_i = \frac{(\hat{y}_{(i)} - \hat{y})^T(\hat{y}_{(i)} - \hat{y})}{pM \text{S}_{res}}
\]

where \( \hat{y}_{(i)} \) is obtained when removing the \( i \)th observation from the data set. Points with large \( D_i \) values greatly influences the beta estimates, where \( D_i > 1 \) are considered being large. Therefore, it is recommended to eliminate data points exceeding the cutoff-ratio. It is important to remember that the cutoff-ratio only is a recommendation and also based on the sample size \( n \). However, for larger samples sizes, the specified cutoff-ratio makes more sense.

2.5.2 DFFITS

Similar to Cook’s distance, DFFITS, difference in fit or difference in fitted \( \hat{y} \)’s with and without each data point, also measures the influence of one point and is given by

\[
DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}}
\]

where \( \hat{y}_i \) is the fitted value including the point \( i \) and \( \hat{y}_{(i)} \) is the fitted value excluding the point \( i \). The factor \( S^2_{(i)} \) is the R-student estimate of the mean square residual.
shown in Equation (17) and \( h_{ii} \) represents the diagonal element \( i \) from the hat matrix.

The suggested cutoff-ratio is given by \(|DFFITS_i| > \sqrt{p/n}\). The DFFITS is a measure of the number standard deviations the fitted value changes if the \( i \)th observation is removed from the model. It essentially a useful measure when detecting both leverage and prediction error.

### 2.6 Variable Selection

Oftentimes, there are several regressor variables to choose from, making it difficult to pinpoint the likely few important ones. Variable selection is the procedure of finding an appropriate subset of regressor for the model. Further, variable selection is the most common procedure for solving the problems of multicollinearity.\(^{26}\)

#### 2.6.1 Null Hypothesis

As mentioned in Section 2.3.4 the normality assumptions open up the possibility of using hypothesis testing and generating reliable confidence and prediction intervals. The standard test uses the following hypotheses

\[
H_0 : \beta_1 = \beta_{10} \quad H_1 : \beta_1 \neq \beta_{10}
\]

(21)

where \( \beta_{10} \) is set to 0, and the goal is to conclude whether the regression coefficient \( \beta_1 \) is non-zero. Since the errors are normally distributed with expected value zero, the observations are also normally distributed with expected value \( X\beta \), since \( E[y] = E[X\beta + \epsilon] = E[X\beta] + E[\epsilon] = X\beta \). If \( H_0 : \beta_1 = \beta_{10} = 0 \) cannot be rejected the linear relationship between the predictor variable \( x \) and the response \( y \) either does not exist, is not important for explaining the response or does not have a linear relationship with the response.

\(^{26}\)Montgomery, Peck, and Vining, *Introduction to Linear Regression Analysis*. pp. 328
2.6.2 t-test

The t-statistic is the ratio of the departure of the estimated value of a parameter from its hypothesized value to its standard error. By assuming that Y is a random variable following a normal distribution with mean \( \mu \) and variance \( \sigma^2 \) we have \( Y \sim \mathcal{N}(\mu, \sigma^2) \to Z = \frac{Y - \mu}{\sigma} \sim \mathcal{N}(0, 1) \), implying that \( Z^2 \) (\( Z \) being the standard normal variable with 1 degree of freedom) follows a \( \chi^2 \) distribution, i.e. \( Z^2 \sim \chi^2_1 \). Thus, the square of a standard normal variable is a \( \chi^2 \) random variable with one degree of freedom. If the null hypothesis given in Equation (21) is true, the \( t \) statistic \( t_0 \) follows a \( t_{n-2} \) distribution.

\[
t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{res}/S_{xx}}} \tag{22}
\]

The degree of freedom for \( t_0 \) is the same as \( MS_{res} \). Using Equation (22) the observed value \( t_0 \) can be compared to the upper \( \alpha/2 \) percentage point of the \( t_{n-2} \) distribution (\( t_{\alpha/2,n-2} \)). The \( t \) statistic for the null hypothesis is given by

\[
t_0 = \frac{\hat{\beta}_1}{\sqrt{MS_{res}/S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \tag{23}
\]

since \( \beta_{10} = 0 \) from the Equation (21). If \( |t_0| > t_{\alpha/2,n-2} \), then the null hypothesis is rejected.

2.6.3 F-test

The F-statistics shows the relationship between two independent variables where both of the variables have the \( \chi^2 \) distribution with different degrees of freedom. For example, if we introduce two variables \( X \) and \( Y \) where \( X \sim \chi^2_v \) and \( Y \sim \chi^2_n \). Thus, the ratio is now given by

\[
\frac{X/v}{Y/n} \sim F_{v,n} \tag{24}
\]

which is the \( F \) distribution with \( v \) and \( n \) degrees of freedom. This shows that the
ratio of two independent random variables that follow a $\chi^2$ distribution, follows an $F$ distribution.

Therefore, the $F$ test can be used to test the null hypothesis. In this case, the null hypothesis is that all of the regression coefficients are equal to zero, implying, if the hypothesis is not rejected, that the model has no predictive capability. If the null hypothesis is true, the following holds

$$\frac{MS_R}{MS_{res}} \sim F_{1,n-2}$$  \hspace{1cm} (25)

since $MS_R = SS_R/1$ has one degree of freedom and $MS_{res} = SS_{res}/(n-2)$ has $(n-2)$ degrees of freedom.

$$F_0 = \frac{MS_R}{MS_{res}}$$  \hspace{1cm} (26)

The null hypothesis is rejected if $F_0 > F_{\alpha,1,n-2}$, where $\alpha$ is the given level of significance.

### 2.6.4 All Possible Regressions

As stated earlier, it is desirable to choose a subset of the candidate regressors. By fitting various combinations of the regressors and comparing them based on certain criteria, models with a certain setup of parameters can be selected.

All possible regressions is one method of achieving this. By fitting the model with the regressors one by one and comparing them with each other and choosing the best ones, based on some criteria, the best regression model can be selected.

Assuming that the intercept $\beta_0$ is included in each model, there is a total of $2^p$ models to evaluate. One obvious drawback with the method is the exponentially growing number of models based on $p$, for an instance $2^{10}$ equals to 1024 models, however modern day computers and software have made the procedure relatively simple to use.
2.6.5 Backward Elimination

Backward Elimination is a step-wise regression method used in situations where all possible regressions can be burdensome computationally. This method begins with the model including all $p$ candidate regressors. The corresponding t-test or F-test is computed for each of the regressor. Comparing these t-tests or F-tests to a pre-selected value, called $t_{\text{leave}}$ for example, the regressor corresponding to the smallest value of the t-test or F-test is removed from the model.

This procedure is repeated $k$ times until the model contains a number of $p - k$ regressors, where each of the corresponding test statistic value is greater than the cutoff value.

2.6.6 Cross Validation

In situations when there is no new fresh data to validate the model, tools as cross validation can be used. The data set is split into two parts, one is the estimation part, and the other is the prediction part. A new regression model is build, based on the estimation data, and then validate the model with the prediction data. If the original data is collected within a time frame, one can use time as the basis when splitting the data set into these two components. However, the basis which can be chosen arbitrarily has a probability of not stressing the model enough, and therefore it would be preferable to use several estimation and prediction-sets to improve the validation technique.

2.6.7 R-squared and Adjusted R-squared

To measure the overall adequacy of the model, one can use the coefficient of determination called "R squared". The proportion of the variance in the dependant variable that is predictable from the independent variable(s) is given by $R^2$. In other words, $R^2$ measures how close the observed data are to the fitted regression line.

The residuals sum of squares ($SS_{\text{res}}$), the discrepancy between the data and the estimated model, divided by the total sum of squares ($SS_{\text{tot}}$), the squared differences
of each observation from the overall mean, minus 1 gives us R-squared.

\[ R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = 1 - \frac{SS_{res}}{SS_{tot}} \in [0, 1] \quad (27) \]

However, the \( R^2 \)-measure has some limitations, one of them being that it increases when more predictors are added to the model. The modified version of the measurement, the adjusted R squared \( R^2_{adj} \), only increases when the added predictors improves the model.

\[ R^2_{adj} = 1 - \frac{SS_{res}/(n - p)}{SS_{tot}(n - 1)} \quad (28) \]

Thus, high R-values are desirable when measuring the overall model adequacy.

### 2.6.8 Mallow’s Cp Statistic

Mallow’s \( C_p \) addresses the issue of overfitting, i.e. when a model contains more parameters than can be justified by the data. The statistic is

\[ C_p = \frac{SS_{res}}{MS_{res}} - n + 2p \quad (29) \]

where \( n \) denotes the number of observations and \( p \) the number of parameters. By plotting \( C_p \) against \( p \), the models showing lowest \( C_p \)-values and which also are closest to the line \( C_p = p \) are the best. Figure 2.2 displays a graph of the statistic and indicates that model C, with four parameters, although above the line \( C_p = p \) compared to model A, is better than A, since it is below A, thus representing a model with lower total error.
Mallow’s $C_p$ can be used as one of the criteria in the all possible regression procedure for checking the model adequacy.

2.6.9 Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

Originating from the AIC, BIC is simply an extension of AIC. Acting as criteria, for instance in the all possible regression procedure, these criteria help to choose the best predictors by penalizing the addition of regressors as the sample size increases.

Since some information almost always will be lost during statistical modeling when using the model to represent the process, the models with the least loss of information are preferred. AIC measures the entropy of the model, i.e. the expected information of the model, were the lowest AIC values are most desirable and BIC measures the posterior probability of a model being true, meaning similar to AIC that lower BIC-values are preferred. The most significant difference between these criteria is their size of the penalty — BIC penalizes model complexity more heavily.

Simply put, AIC and BIC consist of a ”measure of fit” and a ”complexity penalty part”. The likelihood function $L$, the plausibility of a value for the parameter given some data, represents the goodness of fit while $p$, the number of parameters, rep-
resents the complexity.

\[
AIC = -2 \ln (L) + 2p = n \ln \left( \frac{SS_{res}}{n} \right) + 2p
\]  

(30)

\[
BIC = -2 \ln (L) + p \ln n = n \ln \left( \frac{SS_{res}}{n} \right) + p \ln n
\]  

(31)

Summarized, these criteria ask whether more regressors should be added to the model, since \(SS_{res}\) cannot increase when the number of regressors is increased, there exists a trade-off between the goodness of the fit of the model and the complexity of the model.

### 2.7 Transformations

When starting the regression analysis, the usual assumption about linearity is often made. With this assumption, the following needs to be satisfied:

1. The model errors have mean zero, constant variance and uncorrelated.
2. The model errors follows a normal distribution.
3. The form of the model is correct.

If these assumptions are incorrect, a transformation of the data needs to be done in order to find a relationship between the regressor and the regressor variables.

#### 2.7.1 Box-Cox

One such method is called the Box-Cox method. This method uses maximum likelihood to find the power of \(\lambda\) to use in the transformation \(y^{(\lambda)}\). Because of the discontinuity when \(\lambda\) approaches zero, the equations are given by:

\[
y^{(\lambda)} = \begin{cases} 
\frac{y^{\lambda-1}}{\lambda}, & \lambda \neq 0; \\
\ln y, & \lambda = 0; 
\end{cases}
\]  

(32)
Where \( z = \ln^{-1}[1/n \sum_{i=1}^{n} \ln y_i] \) is the geometric mean value of the observations \( y_i \).

The Jacobian of the transformation and the divisor \( z^{\lambda-1} \) are related when converting \( y \) into \( y^{(\lambda)} \) which ensures the residual sum of squares with different values of the power \( \lambda \) to be comparable.
3 Data

By choosing the most popular stock and volatility indices, alongside well established EPU indices, the idea was to include a mix of local and international measures on price and volatility changes and financial, political and economic uncertainty. The equity and volatility indices were accessed through the Thomson Reuters Eikon software. The data was transferred in an \texttt{xlsx} format into the programming language \texttt{R}. The relevant information, such as Exchange Date and Close, was extracted and aggregated.

Since indices put simply are, for example, aggregated stock prices, the relevant information is whether or not an index has increased, since the nominal value does not say much. Therefore, the daily returns of the indices are calculated by dividing their respective closing prices $p_n$ for day $n$, with their previous price, i.e. $p_{n-1}$. Subtraction by one presents it percentage.

\[
\text{Daily Return} = \frac{p_n}{p_{n-1}} - 1
\]

Thus, an aggregated data frame consisting of the matching dates and daily returns of the equity and volatility indices and the EPU indices was created.

However, since some of the rows of the date frame consisted of empty values, because of indices not matching opening exchange dates, these rows were omitted from the data frame, i.e. rows that consisted of at least one \texttt{N/A}-element were deleted. Therefore, the final data strongly depends on there being available and consistent daily data for each index.

3.1 Trading Volume

For listed shares the trading volume can be defined as the total number of traded shares over the day, i.e. the total number of all shares that changed hands, and in our case —the total number of all of Avanza’s clients shares that changed hands. Since each trade, generally, results in a commission fee, aggregating the commission revenue for each day produces an index representing the trading volume per
day, i.e. the revenue generated from the commission fees are in theory perfectly correlated with Avanza’s customers activity.

3.2 Equity

Some of the stock indices used are described in this section while all of the variables, including the volatility and EPU indices, can be found in Table 3.1.

The Dow Jones Industrial Average DJI represents the stocks of 30 large and well-known U.S. companies covering all industries with the exception of transportation and utilities. The stocks are selected by editors of The Wall Street Journal. The index is price weighted.

The S&P 500 Index SPX represents over 500 large U.S. companies and captures approximately 80% of available market capitalization.

Alongside SPX and DJI, the NASDAQ Composite Index IXIC is one of the three most followed indices in the U.S. stock market. The index contains data of information technology companies and includes over 3,300 common equities listed on the exchange.

The S&P/TSX Composite Index GSPTSE represents the Canadian benchmark index through about 250 companies and roughly 70 percent market capitalization on the Toronto Stock Exchange.

The OMX Stockholm 30 Index OMXS30 is the Stockholm Stock Exchange’s leading share index. This index consists of the 30 most actively traded stocks on the Stockholm Stock Exchange.

The OMXSSCPI Index consists of all Small Cap companies listed on NASDAQ OMX Stockholm Exchanges. The group of Small Cap companies includes companies whose shares have a market value of less than 150 million euro.
3.3 Volatility

The CBOE Volatility Index $\text{VIX}$ represents the options on the Chicago Board Options Exchange (CBOE), i.e. options betting against S&P 500, that can be used to hedge against volatility spikes. In other words, one could say the index measures public concern —hence the nickname of the index: "the fear index".

The CBOE DJIA Volatility Index $\text{VXD}$, similar to $\text{VIX}$, is based on the prices of options, but unlike $\text{VIX}$, options betting against DJI. Hence, it measures the investors near time expectancy (30-days) of the stock volatility.

The Euro STOXX Volatility Index $\text{V2TX}$ is based on the real time option prices in order to reflect the market expectations of both short-term and long-term volatility. The measurement is constructed by taking the square root of the implied variance across all options of a given time to expiration.

3.4 Economic Policy Uncertainty (EPU)

EPU is an index of the frequency of relevant selected words in several news outlets and there are several EPU indices to choose from country wise. Assuming that investors are interested in following their investments in the respective country, some indices are more relevant than others. For example, as of now, out of the ten most owned stocks by customers between the ages of 18 and 30, seven are Swedish, two are American and one is Canadian. Because of the geographical limitation to the Swedish market, the American and the British indices were chosen since both USA and UK influences the Swedish news.

The American EPU index consists of data from several different newspapers, for example USA Today which is a national paper, but also smaller local ones. The American EPU index is constructed from three primary sources; "economic/economy", "uncertain/uncertainty" and "legislation/deficit/regulation/congress/federal reserve/white house". If any of these three groups of words exist in the paper, the index will rise. There are both monthly and daily indices of the EPU, where the index ordered on a daily basis is chosen.\footnote{Nick Bloom, Scott R. Baker and Steven J. Davis, \textit{US Daily News Index}}

27Nick Bloom, Scott R. Baker and Steven J. Davis, \textit{US Daily News Index}
The British EPU index is used for investigating British policy-related economic uncertainty. Both the American and the British index counts the number of articles containing any of a pre-defined number of words, which for the British index is represented by "policy", "tax", "spending", "regulation", "Bank of England", "budget" and "deficit".\textsuperscript{28}

The Swedish index is constructed in the same way as for the American and British indices,\textsuperscript{29} but because of the lack of data regarding this index, it had to be excluded from further study. The data of the Swedish EPU index could only be generated on a monthly basis, which is the reason why it was excluded since it significantly reduced the number of observations in our study.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commission Revenue</td>
<td>Avanza’s aggregated daily commission revenue</td>
</tr>
<tr>
<td>DJI</td>
<td>Dow Jones Industrial Average, equity index</td>
</tr>
<tr>
<td>SPX</td>
<td>Standard &amp; Poor’s 500, equity index</td>
</tr>
<tr>
<td>OMX</td>
<td>Swedish stock market, equity index</td>
</tr>
<tr>
<td>IXIC</td>
<td>Nasdaq Composite, equity index</td>
</tr>
<tr>
<td>NDX</td>
<td>Nasdaq 100, equity index</td>
</tr>
<tr>
<td>SIX</td>
<td>All listed Swedish companies, equity index</td>
</tr>
<tr>
<td>GSPTSE</td>
<td>Canadian equivalent to the S&amp;P 500, equity index</td>
</tr>
<tr>
<td>VIX</td>
<td>CBOE Volatility Index, volatility index</td>
</tr>
<tr>
<td>VXN</td>
<td>30-day market expectations of the NDX, volatility index</td>
</tr>
<tr>
<td>V2TX</td>
<td>Euro STOXX, volatility index</td>
</tr>
<tr>
<td>VXD</td>
<td>CBOE DJIA, volatility index</td>
</tr>
<tr>
<td>GSPTXVL</td>
<td>Composite High Beta, volatility index</td>
</tr>
<tr>
<td>EPU U.S.</td>
<td>American economic policy uncertainty index</td>
</tr>
<tr>
<td>EPU U.K.</td>
<td>British economic policy uncertainty index</td>
</tr>
<tr>
<td>FRED</td>
<td>Federal Reserved Economic Data, volatility index</td>
</tr>
<tr>
<td>OMXSSCPI</td>
<td>Stockholm Small Capital Price Index, volatility index</td>
</tr>
<tr>
<td>OMXSPI</td>
<td>Stockholm Price Index, volatility index</td>
</tr>
<tr>
<td>TOPX</td>
<td>Tokyo Stock Exchange Price Index, volatility index</td>
</tr>
<tr>
<td>STOXX</td>
<td>Price weighted, equity index</td>
</tr>
</tbody>
</table>

Table 3.1: Description of the regressor variables

\textsuperscript{28} Nick Bloom, Scott R. Baker and Steven J. Davis, \textit{UK Daily News Index}

\textsuperscript{29} Nick Bloom and Davis, \textit{Sweden Monthly EPU Index}
4 Result

First, the full model with \( p = 19 \) parameters and \( n = 1594 \) observations is fitted, producing the first unprocessed linear regression model on the form in Equation (3). Then, a thorough residual analysis is performed were potential outliers are identified and analyzed. Third, the need of a transformation is investigated. If a transformation is performed, a new residual analysis and a potential transformation are once investigated and performed. If the need for a transformation does not exist, a variable selection is conducted and the final models are selected.

4.1 Residual Analysis

After fitting the full model a thorough residual analysis again is performed. Plotting the residuals, information about potential outliers and their behaviour is given. The standardized residuals in Figure 4.1 show that the majority of residuals reside in the acceptable area, i.e. between minus and plus two, while some observations, such as observation 1323, 952 and 1039, do not. These observations outside the red lines are potential outliers and should be further analyzed.

![Studentized Residuals](image)

**Figure 4.1: Studentized residuals plotted against the observations**
Plotting the PRESS-residuals, the observations that are associated with large $h_{ii}$ will usually have large PRESS residuals, and generally be high influence points. Figure 4.2 indicate that the observations with large PRESS-residuals are the same as the ones with large residuals in Figure 4.1.

The observations large PRESS-residual values indicate that they are likely to be observations were the model fits reasonably well, but does not provide good predictions of fresh data. Also, the sum of the PRESS-residuals, displayed in the title as close to $6.53e+26$, is relatively large, which is not desired.

![PRESS residuals plot](image)

**Figure 4.2:** PRESS-residuals plotted against the observations

### 4.2 Leverage and influential points

To extend the analysis of potential outliers and influential observations, both Cook's distance and DFFITS are investigated.

Cook's distance shows three observations with large $D_i$-values, indicating that they influence the beta estimations (see Figure 4.3). However, the observations $D_i$-values do not exceed the recommended cutoff-value of $D_i > 1$. Therefore, Cook's recommends us to not delete them.
Compared to Cook’s distance, DFFITS displays several observations that are well above the recommended threshold, the most extremes ones being 952 and 1196 (see Figure 4.4). DFFITS indicates that these observations should be deleted given the $\sqrt{p/n}$ threshold.

Figure 4.4: DFFITS plotted against the observations
Since the deletion diagnostics Cook’s and DFFITS do not give the same recommendation, another type of diagnostic is performed. The outlier and leverage diagnostics are displayed in Figure 4.5 and show observations categorized as outliers and (or) as having leverage. Clearly, the non-outlier observations that only provide leverage are desirable for the estimation procedure, while the outlier observations that contribute with leverage are not, and should be removed. However, the outlier and leverage observations cannot be removed without an explanation, i.e. are they bad data or just unusual perfectly explainable events?

![Outlier and Leverage Diagnostics for Courtage](image)

**Figure 4.5:** All the observations classified as normal, leverage, outlier or outlier & leverage observations

The most significant outliers are gathered in Table 4.1 and analyzed more in detail. Connecting the observations to their respective dates and commission revenues, reasonable explanations quickly add up.
<table>
<thead>
<tr>
<th>Observation #</th>
<th>Date</th>
<th>Δ Comission revenue</th>
<th>Plausible explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>952</td>
<td>2016-06-27</td>
<td>110.60%</td>
<td>Brexit</td>
</tr>
<tr>
<td>1039</td>
<td>2016-11-09</td>
<td>56.05%</td>
<td>Trump</td>
</tr>
<tr>
<td>1322</td>
<td>2018-02-05</td>
<td>30.21%</td>
<td>Unexpected inflation</td>
</tr>
<tr>
<td>1323</td>
<td>2018-02-06</td>
<td>39.56%</td>
<td>Unexpected inflation</td>
</tr>
</tbody>
</table>

Table 4.1: The most severe outliers and the date of their occurrences

Starting with outlier 952 and the events on its date 2016-06-27, Avanza’s customers broke a record in the number of stock transactions during one day, with Brexit acting as a catalyst.\(^{30}\) Secondly, regarding observation 1039 and on its date 2016-11-09, Trump was elected as the president of the U.S. —resulting in increased uncertainty and an overall decreased willingness to invest, increasing Avanza’s commission revenue with over 50 percent.\(^{31}\) Finally, regarding the two last observations, on both their dates 2018-02-05 and 2018-02-06, the largest downfall on the stock exchange after BREXIT, took place, apparently caused by a large downfall after unexpected high salaries numbers in the U.S. was presented —indicating that the inflation might rise sooner than expected.\(^{32}\)

In summary, these outliers are clearly unusual, however they are fully plausible observations and should therefore not be removed from the data set since they are not bad observations such as faulty measurements or incorrect recording of data.

### 4.3 Transformations

After fitting the full model and conducting a thorough residual analysis, the need for a transformation is investigated. By plotting the log-likelihood as a function of \(\lambda\) with a 95% interval, Figure 4.6a show that \(\lambda\) is outside of the interval, indicating a need for a transformation.

\(^{30}\)SvD, *Rekordhandel på börsen efter brexit*

\(^{31}\)Privata Affärer, *Experterna spår kraftig börsnedgång*

\(^{32}\)SvD, *Största börstappet sedan Brexitomröstningen*
The value chosen for $\lambda$ is 0.5 because of its appropriateness. After conducting the same previous residuals analysis as before, the same outliers as in Table 4.1 are identified by the residuals, Cook’s distance, DFFITS and the outlier and leverage diagnostics. Further, after performing the box-cox method again, no indication of a transformation is found, since $\lambda = 1$ is found in the 95% interval (see figure 4.6b).

### 4.4 Variable Selection

After identifying the outliers and performing eventual transformations, the task of producing a final model still exists. Through variable selection some final model(s) are produced.

#### 4.4.1 All Possible Regressions

Proceeding with the all possible regressions analysis, criterion such $R$-squared, adjusted $R$-squared, Mallow’s $C_p$ and Schwartz’s BIC are chosen. The criterion are displayed in Figure 4.6 were the black boxes on the $x$-axis display whether or not the specific regressor variable is included for that specific model on the $y$-axis. For example, observing the Mallow’s $C_p$ plot, the model with the value $C_p = 8$ consists of the variables DJI, SPX, IXIC, VIX, V2TX and OMXSCCPI (excluding the
intercept), while the model with the value $C_p = 17$ only consists of the variables DJI, SPX and V2TX.

![R-squared vs Adjusted R-squared](image1)

![Mallows Cp vs Schwartzs Bayesian Information Criterion](image2)

Figure 4.6: All possible regressions, from top left to bottom right: $R^2$, $R^2_{adj}$, $C_p$ & BIC

The first impression is that several of the regressor variables are left out in many of the models. For example OMX and both the EPU indices. On the other hand, some variables seems to be dominant throughout the criteria such as DJI, SPX, IXIC and V2TX.

Judging by the plots, the all possible regression procedure recommends a model with 10 variables for maximizing $R^2$ and $R^2_{adj}$ values. For minimizing Mallow’s $C_p$ it recommends a model with 9 variables, and for minimizing BIC, a model with
only 1 variable.

4.4.2 Backward Elimination

Combining the all possible regressions method of choosing variables with the backward elimination method, a model can be chosen. Remember that backward elimination omits variables based on their $t$ and $F$ statistics. Backward elimination creates a model with 9 variables, the same as the model minimizing the Mallow’s $C_p$.

<table>
<thead>
<tr>
<th>Intercept</th>
<th>DJI</th>
<th>SPX</th>
<th>IXIC</th>
<th>SIX</th>
<th>VIX</th>
<th>VzTX</th>
<th>GSPTXLV</th>
<th>OMXSPI</th>
<th>TOPX</th>
</tr>
</thead>
<tbody>
<tr>
<td>1286.4</td>
<td>11071.1</td>
<td>-13964.0</td>
<td>4664.9</td>
<td>-12151.6</td>
<td>246.9</td>
<td>174.3</td>
<td>-2097.5</td>
<td>11906.1</td>
<td>-724.2</td>
</tr>
</tbody>
</table>

Table 4.2: 9 variable model presented by backward elimination

However, by comparing the backward elimination model (see Table 4.2) with, perhaps, a model with lesser variables, the backward elimination model seems somewhat over exaggerated. This is because the all possible regressions method show that almost the same $R^2$ and $R^2_{adj}$ values can be achieved by omitting some variables. For example, by omitting two variables (OMXSPI and SIX) from the backward elimination model with nine parameters, the $R^2_{adj}$ only decreases with $\Delta R^2_{adj} \approx -0.00021$. This is a relatively cheap trade-off for a less complex model.

4.4.3 Cross Validation

Having learned that removing some variables from the backward elimination model does not significantly damage the models overall performance, another useful method for choosing variables is used.

The cross validation with $k = 10$ folds minimizes the root mean square error (RMSE) and mean absolute error (MAE). Observing Figure 4.7, the method chooses a model with five variables. Compared to the models containing more or less variables, the model with five variables clearly differentiate itself from the other models. This model is desirable because of its low complexity and errors.
The five variables are and their estimated coefficients are

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>DJI</th>
<th>SPX</th>
<th>IXIC</th>
<th>V2TX</th>
<th>OMXSSCPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>1285.332</td>
<td>11285.454</td>
<td>-15370.214</td>
<td>3715.533</td>
<td>200.584</td>
<td>-1871.988</td>
</tr>
</tbody>
</table>

Table 4.3: Final chosen model, proposed by 10-folded cross validation

In conclusion, the model chosen by ten-fold cross validation only differs $\Delta R^2_{adj} \approx -0.00262$ from the model chosen by all possible regressions $R^2$ and $R^2_{adj}$, while consisting of four less variables and lower mean error (compare index 5 and 9 in Figure 4.7).

### 4.4.4 OLS-assumptions

Proceeding with the final model presented in Table 4.3, the OLS-assumptions are checked.

As stated before, OLS assumes that the random errors have a mean of zero. Calculating the mean of the fitted models residuals indicate a mean value near zero ($\text{meanValue} = -3.424928\times10^{-14} \approx 0$), indicating strict exogenity.
Homoscedasticity is controlled by plotting the residuals of the fitted model against its fitted values (see Figure 4.8). The red line in the figure estimates the spread of the observations. The majority of the observations seem to lie on a horizontal band, which is positive, however, the red line is clearly bent upwards, both in its tail and head, indicating heteroscedasticity. The outliers with leverage (see Table 2.3) are probably the cause of this.

![Plot of residuals versus fitted values](image)

**Figure 4.8:** Plot of residuals versus fitted values

Further, OLS assumes that no auto-correlation should exist when estimating the beta coefficients. The detection of auto-correlation can be displayed by plotting the residuals against the time. Observing Figure 4.9 some kind of correlation seems to exist, whether it is negative or positive correlation is hard to decide. But since the residuals do not alternate signs rapidly in comparison to residuals with similar sign appearing in clusters, they are probably positively correlated. There also seems to exist an overall trend were the residuals increase with time. The conclusion is nevertheless same, auto-correlation is detected and the assumption is violated.
The assumption about a normal distribution is checked by observing Figure 4.10, the grey area in the background with a wider "tail" and "head" is the acceptable range for the sample to fulfill the normality assumption. Clearly, both the errors tail and head are outside of the grey area, resembling a heavy-tailed distribution. Thus, it is concluded that the normality assumptions are not fulfilled.
Multicollinearity can be detected through the calculation of the VIF:s and the eigenvalues. The VIF:s (see Table 4.4) exceed the recommended threshold values 5 and 10, indicating a case of multicollinearity for the choice of variables. The eigenvalues, as well as the condition number, are also calculated. The condition number well exceeds the recommended threshold 100 ($\kappa = 788.619$) and is close to threshold indicating severe multicollinearity, i.e. $\kappa = 1000$. The condition number confirms the VIF:s analysis of a case of severe multicollinearity.

<table>
<thead>
<tr>
<th></th>
<th>DJI</th>
<th>SPX</th>
<th>IXIC</th>
<th>V2TX</th>
<th>OMXSSCPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>18.985775</td>
<td>41.715249</td>
<td>11.744280</td>
<td>1.490258</td>
<td>1.371106</td>
</tr>
</tbody>
</table>

Table 4.4: Variance inflation factors

The OLS-assumptions, except for multicollinearity, are summarized in Table 4.5.
<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>p-value</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Stat</td>
<td>109.184</td>
<td>0.000e+00</td>
<td>Assumptions NOT satisfied!</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.434</td>
<td>2.312e-01</td>
<td>Assumptions acceptable.</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>18.765</td>
<td>1.479e-05</td>
<td>Assumptions NOT satisfied!</td>
</tr>
<tr>
<td>Link Function</td>
<td>31.019</td>
<td>2.555e-08</td>
<td>Assumptions NOT satisfied!</td>
</tr>
<tr>
<td>Heteroscedasticity</td>
<td>57.966</td>
<td>2.665e-14</td>
<td>Assumptions NOT satisfied!</td>
</tr>
</tbody>
</table>

Table 4.5: OLS assumptions, level of significance = 0.05

Global Stat shows the relationship between the predictors and response and indicates a non-linear relationship between at least one of the predictors versus the response, since the p-value rejects the null hypothesis, i.e. $p < 0.05$. The Skewness is accepted, indicating that the distribution is neither positively nor negatively skewed and that there is no need for transformation. Kurtosis is not satisfied, indicating that the assumption of normality does not hold given the level of significance. A rejected Link Function indicates that an alternative form of the generalized linear model, such as logistic or binomial, should be used. Last, Heteroscedasticity is the opposite of homoscedasticity, indicating that the errors do not have a constant variance for each observation.
4.5 Final Model

In the end the best model, given the circumstances, is a model consisting of 5 variables with equity and volatility indices and no EPU indices on the form

\[ \sqrt{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_5 x_5 \]  

(34)

More specifically the model consists of the coefficients in Table 4.6 leading to the model on form (34) to be written as

\[ \sqrt{\text{Daily aggregated commission revenue}} = 1286.4 + 11112 \times \text{DJI} \]
\[ -14356 \times \text{SPX} + 4899.9 \times \text{IXIC} - 189.31 \times \text{V2TX} - 2093.6 \times \text{OMXSSCPI} \]  

(35)

The partial residual plots are displayed in Figure 4.11 and the estimates of the coefficients and their limits in a 95 percent interval, alongside their p-values, are presented in the Table 4.6
Figure 4.11: Partial residuals plots

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Lower limit (5.0%)</th>
<th>Upper limit (95%)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1285.903</td>
<td>1275.029</td>
<td>1296.777</td>
<td>2e-16</td>
</tr>
<tr>
<td>DJI</td>
<td>11806.870</td>
<td>6177.023</td>
<td>17436.72</td>
<td>4.11e-05</td>
</tr>
<tr>
<td>SPX</td>
<td>-16824.578</td>
<td>-25022.71</td>
<td>-8626.443</td>
<td>5.98e-05</td>
</tr>
<tr>
<td>IXIC</td>
<td>4634.423</td>
<td>996.5248</td>
<td>8272.321</td>
<td>0.0126</td>
</tr>
<tr>
<td>V2TX</td>
<td>221.609</td>
<td>41.98532</td>
<td>401.2327</td>
<td>0.0157</td>
</tr>
<tr>
<td>OMXSSCPI</td>
<td>-1246.238</td>
<td>-2779.493</td>
<td>287.0169</td>
<td>0.1112</td>
</tr>
</tbody>
</table>

Table 4.6: Coefficient confidence interval
5 Discussion

5.1 Methodology

As displayed in Table 4.6, the two most significant variables in the final model are DJI and SPX. Since both are most certainly the largest and most popular stock indices used within the financial market, it seems reasonable that they would capture some of the explanation, such as Great Britain leaving the E.U. or Trump being elected as president.

Apart from the three equity indices, two volatility indices were included in the final model. Although included, V2TX and OMXSCPI were not as significant as the equity indices and further, more established volatility indices, such as the VIX, was entirely excluded from the model (see Figure 4.6). This seems odd, since in general, periods of increased volatility imply higher uncertainty and therefore a tendency to invest less, while short periods of volatility spikes induces higher trading activity. The latter, we believe, is somewhat shown in Avanza’s customers trading behaviour. Based on the outlier analysis in Section 4.2, Avanza’s customers trading activity spiked during these unexpected events, as we would have predicted. Again however, none of the volatility nor uncertainty indices were significant in explaining these.

One possible explanation for the insignificance of the volatility and uncertainty indices is lag. For example, extracting the four dates 2018-02-02-05-06-07 and their respective returns and commission revenue values, some observations can be made (see Table 5.1). The U.S. EPU index more than doubles from the first to the second observation, while VIX and the commission revenue also increase. This seems reasonable —a relevant event causes increased uncertainty and volatility, which is noticed by Avanza’s customers, hence increasing their transaction activity and Avanza’s aggregated commission revenue. However, the following day the U.S. EPU index falls substantially, while the transaction activity and VIX continues to increase, i.e. now the U.S. EPU index and Avanza’s commissions revenue are negatively correlated as opposed to the day before.

One theory is that right at the moment when an unexpected event occurs both the
uncertainty indices and the investors trading activity correlate, but not after. The
day after the public learned about the unexpected inflation they maybe were not
so uncertain anymore, decreasing the EPU index, while the investors still traded,
causing a lag.

<table>
<thead>
<tr>
<th>Observation #</th>
<th>Date</th>
<th>Δ Commission revenue</th>
<th>VIX</th>
<th>EPU U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1321</td>
<td>2018-02-02</td>
<td>8.41%</td>
<td>18.6%</td>
<td>22.9%</td>
</tr>
<tr>
<td>1322</td>
<td>2018-02-05</td>
<td>30.21%</td>
<td>32.2%</td>
<td>138%</td>
</tr>
<tr>
<td>1323</td>
<td>2018-02-06</td>
<td>39.56%</td>
<td>10.7%</td>
<td>-36.5%</td>
</tr>
<tr>
<td>1324</td>
<td>2018-02-07</td>
<td>-43.25%</td>
<td>-15.2%</td>
<td>-37.0%</td>
</tr>
</tbody>
</table>

Table 5.1: The most severe outliers and the time of their occurrences

By our definition, unexpected means that the majority was not able to foresee it.
Hence, the surprised investors will try to trade as much as possible, as soon as
possible, to minimize their losses, while other investors might try to maximize
their returns given the time frame —increasing the overall trading activity during,
but also shortly after, the initial event has taken place.

On the other hand, the EPU index can also be flawed. Since it consists of key
words aggregated from major news outputs spread across a certain geographic
area, it depends on what gets written, and how much is written about it. Usually
the news will bombard with output weeks after special occurrences, potentially
increasing the uncertainty index although the uncertainty is gone.

In summary, there might exist a time delay between the unexpected event and the
investors activity. Further, the investors activity might continue for some time
after the event has become known to the public, which the uncertainty indices do
not reflect, while the commission revenue does. And vice versa, newspapers might
write as much about events one week later as the day it happened, thus displaying
misleading correlation. But this still does not explain why the volatility index, that
is not built by the same method as the EPU indices, is not significant. Maybe the
adequacy of our model can act as a source of explanation.
5.2 Model Adequacy

The final model's adjusted R-squared value is 0.01926, i.e. this model explains approximately 1.926% of the observed variance. The question of whether this is a good or bad value is irrelevant, since it depends on how the variables are measured, whether any transformations have been done etc. and whether the model can actually be trusted.

Avanza's daily commission revenue is a time series, i.e. it is a series of data points indexed in time order. When the response variable is a time series, much of the response variable's predictive power is often a result of its own history via lags, differences and/or seasonal adjustments, which clearly seems to be the case here as well (see Figure 4.9). Therefore, it is unwise to evaluate its predictive power when we do not know what lies behind it.

Further, many of the OLS-assumptions are violated (see Figure 4.5). These assumptions are vital for obtaining accurate estimations of the OLS beta coefficients. The presence of multicollinearity, a non-normal distribution, heteroscedasticity and kurtosis all contribute to less accurate and misleading beta estimations.

In summary, the model's poor adequacy could be a potential source of explanation when it comes to explaining certain factors, that in theory are significant for explaining the trading volume, were ruled out as insignificant.

5.3 Previous studies

In general, there are numerous factors affecting individual trading behaviour. As stated in the introduction, this has been shown by studies were factors such as price changes, volatility and even uncertainty play a significant role. By that, it seems as if the factors we have used in this study may be miss used, rather than irrelevant, with regards to the investigation of their relationship with the response variable.

One important distinction is the type of method that we have used and the population that was studied. First, regression modeling is a powerful and frequently used statistical tool, especially within finance subjects, but almost no studies regarding
our subject using regression analysis were found. One reason might again be that a time series analysis should be used instead of conducting a regression analysis, mainly because of the lag and auto correlation. Secondly, usually when referring to price and volume, the meaning of the concepts are the stock and its price and volume. This report refers to several stock indices and Avanza’s aggregated commission revenue. These two relationships are of course different, which is why the established price-volume correlation between a stock’s price and volume should not be easily accepted here.

In conclusion, the studies found do neither strengthen nor weaken the results, actually not much research about the fluctuating revenue streams of online stock brokers and the cause of them has been found at all. However, the literature that is used is still valuable for giving some insight into the mechanisms of investors trading behaviour.

5.4 Future Work

There are several suggestions where to start (regarding) future work. For example, by using generalized linear models (GLM) one could surpass the assumptions of no auto correlation and a normal distribution, since GLM does not assumes that the errors are not auto correlated and can choose witch distribution to account for in its modeling.

Moreover, a time series analysis would be recommended in order to analyze the data with a more scientific and statistic approach to improve the reliability of the outcome, since the data is taken at different time periods. The result given in Figure 4.9 showed that the there is a possible trend in the residuals, which must be accounted for.

5.5 Avanza

Because of the poor model adequacy, the results cannot help to explain the overall trading behaviour of their customers. The analyze of the competitiveness through Porter’s five forces can therefore not be used in further implementation for Avanza.
Although some of their customer’s behaviour is known, such as taking the chance to trade directly after unexpected events as when the Swedish bank Swedbank was caught laundering money which led to an 89% increase of Avanza customers buying Swedbank's stock,\(^{33}\) not much fruitful comes out of it. Since these events usually are unexpected, nobody can predict them.

Relevant information for Avanza might be information explaining both long and short term behaviour of their customers. Perhaps, more in detail on how the customers are affected by the seasons of the year, or how trading differs on Mondays and Fridays etc. This could lead to key insights such as, for example, learning that customers are more prone to invest in stocks during springs than autumns, or that customers tend to save more of their salaries directly after receiving it, rather than long after they have received. Information such as this could be used as pillars when developing new products, or taken into account when streamlining, or increasing the money their customers save and perhaps used for forecasting future earnings.

\(^{33}\)Lans, *Kraftig ökning av Swedbankaktieägare*
6 Conclusion

In this study three main research questions were asked.

1. Does uncertainty, in a financial, political and macroeconomic sense, along with market indicators such as stock prices and volatility indices, impact Avanza’s customers trading activity?

2. What are the possible relationships and what do they look like?

3. How should the possible relationships be implemented?

Question two is answered in the final model where the possible relationships are represented through a multiple linear regression model, and question three is answered in Section 5.5.

The answer to question one is: "we cannot say". The hypothesis was that influential exogenous market factors such as political, economical and financial uncertainty, stock price returns and volatility could help to explain Avanza’s customers trading activity. However, scientifically the final model does not fulfill the basic requirements for it to be reliable, hence the significance of its variables with respect to the response variable, cannot be strengthened nor weakened.
References


[18] Definition of “regression” from the cambridge advanced learner’s dictionary & thesaurus ©cambridge university press.


Appendices

List of Figures

1.1 Plot of OMX against months for the time period 2000-2017 .................. 8
2.1 Left: leverage point. Right: influential observation .............................. 17
2.2 $A_C$ plot (from Montgomery, Peck, and Vining (Introduction to Linear Regres-
sion Analysis) pp. 336) ........................................................................ 24
4.1 Studentized residuals plotted against the observations ......................... 31
4.2 PRESS-residuals plotted against the observations ............................... 32
4.3 Cook’s distance plotted against the observations ......................... 33
4.4 DFFITS plotted against the observations ........................................ 33
4.5 All the observations classified as normal, leverage, outlier or outlier & leverage
observations ...................................................................................... 34
4.6 All possible regressions, from top left to bottom right: $R^2$, $R^2_{adj}$, $C_P$ & BIC . 37
4.7 10-fold cross validation ...................................................................... 39
4.8 Plot of residuals versus fitted values .............................................. 40
4.9 Auto-correlation plot ....................................................................... 41
4.10 Normality QQ-plot for checking normality distribution ................. 42
4.11 Partial residuals plots .................................................................. 45

List of Tables

3.1 Description of the regressor variables ................................................. 30
4.1 The most severe outliers and the date of their occurrences ............... 35
4.2 9 variable model presented by backward elimination ....................... 38
4.3 Final chosen model, proposed by 10-folded cross validation ............... 39
4.4 Variance inflation factors ................................................................. 42
4.5 OLS assumptions, level of significance $= 0.05$ ............................... 43
4.6 Coefficient confidence interval ......................................................... 45
5.1 The most severe outliers and the time of their occurrences .............. 47