Factors that influence condominium pricing in Stockholm: A regression analysis

ELIN DAI
LARA GÜLERYÜZ
Factors that influence condominium pricing in Stockholm

A regression analysis

ELIN DAI
LARA GÜLERYÜZ
Acknowledgments

We want to begin by thanking our supervisor Per Jörgen Säve-Söderbergh at the Royal Institute of Technology, Stockholm, Sweden. We want to thank Booli Search Technologies AB for supplying us with data through their API, which made this study possible. A special thank you to Mikaela Lagercrantz, developer at Booli, for the communication and assistance we had via e-mail, and Emil Kaiser, founder and executive vice president at Booli, for inspiring us with the python coding. We would like to thank the students in the class of I-16 for providing us with help and feedback continuously throughout the process. Lastly, a big thank you to each other, being us authors of this project, for always cooperating and working constructively with the goal to achieve results as a team.
Abstract

This thesis aims to examine which factors that are of significance when forecasting the selling price of condominiums in Stockholm city. Through the use of multiple linear regression, response variable transformation, and a multitude of methods for refining the model fit, a conclusive, out of sample validated model with a confidence level of 95% was obtained. To conduct the statistical methods, the software R was used.

This study is limited to the districts of inner city Stockholm with the postal codes 112-118, and the final model can only be applied to this area as the postal codes are included as regressors in the model. The time period in which the selling price was analyzed varied between January 2014 and April 2019, in which the volatility of the time value of money has not been taken into account for the time period. The final model included the following variables as the ones having an impact on the selling price: floor, living area, monthly fee, construction year, district of the city.
Sammanfattning

Denna studie ämnar till att undersöka vilka faktorer som är av betydelse när syftet är att förutsäga prissättningen på bostadsrätter i Stockholms innerstad. Genom att använda multipel linjär regression, transformation av responsvariabeln, samt en mängd olika metoder för att förfina modellen, togs en slutgiltig, out of sample-validerad modell med ett 95%-konfidensintervall fram. För att genomföra de statistiska metoderna användes programmet R.

Table of contents

List of abbreviations and symbols ........................................... 3

1. Introduction ........................................................................ 4
   1.1 Background .................................................................. 4
   1.2 Previous work .......................................................... 4
   1.3 Objective and Aim ...................................................... 5
   1.4 Scope ....................................................................... 5
   1.5 Tools ......................................................................... 6
   1.6 Problem formulation .................................................. 7

2. Theoretical framework and modelling theory ...................... 7
   2.1 An Introduction to multiple linear regression ................. 7
      2.1.1 Necessary assumptions ........................................ 8
      2.1.2 Homoscedasticity and heteroscedasticity ............... 9
         2.1.2.1 Homoscedasticity ........................................ 9
         2.1.2.2 Heteroscedasticity ...................................... 9
         2.1.2.3 Detecting heteroscedasticity ......................... 9
      2.1.3 Ordinary least squares estimation ......................... 10
      2.1.4 Dummy variables .............................................. 11
   2.2. Diagnostics of outliers and influential observations ........ 12
      2.2.1. DFFITS ......................................................... 12
      2.2.2. Covariance ratio ............................................. 13
   2.3 Model transformation .................................................. 13
      2.3.1 Box-Cox transformation ..................................... 13
   2.4 Multicollinearity ......................................................... 14
      2.4.1. Detecting multicollinearity ............................... 15
         2.4.1.1. Variance inflation factors ......................... 15
         2.4.1.2. Linear dependance ................................. 15
         2.4.1.3. Remedies to multicollinearity .................. 15
   2.5 Variable selection methods .......................................... 16
      2.5.1. All possible regressions ................................... 17
      2.5.2. Adjusted $R^2$ ............................................... 17
      2.5.3. Bayesian information criterion ......................... 17
   2.6. Model transformation and validation ......................... 18
      2.6.1. Hypothesis testing .......................................... 18
      2.6.2. p-value ....................................................... 19
      2.6.3. Out of sample validation ................................... 19
3. Method and model development

3.1. Initial data set and preprocessing ........................................... 20
3.2. Variable selection ................................................................. 21
3.3. Model validation ................................................................. 26

4. Results

4.1. The final model ................................................................. 26
4.2. Comparisons with previous work ......................................... 27

5. Analysis and Discussion

5.1. Box-Cox transformation ....................................................... 28
5.2. Handling of outliers influential observations ............................. 29
   5.2.1. Residual analysis .......................................................... 29
          5.2.1.1. Regular residuals ............................................... 29
          5.2.1.2. Studentized residuals ....................................... 30
          5.2.1.3. Standardized residuals ..................................... 31
   5.2.2. Covariance ratio and DFFITS ....................................... 32
5.3. Variable selection ............................................................. 33
   5.3.1. Handling multicollinearity ............................................ 33
   5.3.2. p-value, All possible regression, BIC .............................. 34
5.4. Model validation ............................................................... 37
   5.4.1. Out of sample validation .............................................. 37
   5.4.2. Checking the necessary assumptions ................................. 38
          5.4.2.1. Individual linear dependency ............................... 38
          5.4.2.2. Error mean, variance, normality and correlation ......... 40

6. Conclusion

7. Bibliography
List of abbreviations and symbols

AIC \(\text{Akaike Information Criterion}\)
API \(\text{Application Programming Interface}\)
BIC \(\text{Bayesian Information Criterion}\)
Cov \(\text{Covariance}\)
Covratio \(\text{Covariance ratio}\)
DFFITS \(\text{Difference in fits}\)
E \(\text{Expected value}\)
MS \(\text{Mean Square}\)
OLS \(\text{Ordinary Least Squares}\)
Res \(\text{Residual}\)
RSS \(\text{Residual Sum of Squares}\)
Var \(\text{Variance}\)
VIF \(\text{Variance Inflation Factor}\)

\(X\) \(\text{matrix containing the regressors}\)
\(X^T\) \(\text{transpose of the regressor matrix } X\)
\(X^{-1}\) \(\text{inverse of the regressor matrix } X\)
\(Y\) \(\text{vector containing the response variables}\)
\(y\) \(\text{response variable}\)
\(\beta\) \(\text{vector containing the regression coefficients}\)
\(\hat{\beta}\) \(\text{OLS estimated vector of } \beta\)
\(\hat{y}\) \(\text{vector of fitted values}\)
\(\sigma\) \(\text{the standard error}\)
\(e\) \(\text{error vector}\)
\(p\) \(\text{the number of parameters in the model}\)
\(k\) \(\text{number of regressors}\)
\(n\) \(\text{the number of data points}\)
\(H_0\) \(\text{null hypothesis}\)
\(H_1\) \(\text{alternative hypothesis}\)
\(\alpha\) \(\text{significance value/alpha value}\)
\(s\) \(\text{estimated standard error}\)
\(h\) \(\text{leverage}\)
1. Introduction

1.1 Background

The price development of the housing market is of great importance and significance for the private economy in households and the national economy in general. The periodical increase and decrease of the index representing the estate pricing in Stockholm can be linked to various factors, sometimes to significantly larger factors such as the financial crisis 2008, sometimes to higher/lower interest rates, to different socioeconomic factors, and sometimes to no considerably meaningful factors at all. The market trend and price development analyzed in this case can be said to depend on internal factors. In this thesis the internal factors consist of factors that can be seen as a part of the condominium and market itself, for example how big the condominium is, the district in which the condominium lies in, how old the condominium is and in what season the condominium was sold. There might possibly be a codependence between different factors that affect the pricing in their turn, these will be taken into consideration with various different regression models and theories to help identify possible inter-relations and what consequences they might have on the final model and conclusion.

1.2 Previous work

Alongside the urbanisation in Sweden – both present and past – there has been an increasing need of housing in urban areas; specifically condominiums, to accommodate the growing population. Since the 1990s, the Swedish housing market has been facing a serious housing shortage because of the lack of new construction, which is in itself due to a variety of factors, the most important ones being the rising renovation needs, escalating affordability issues, and municipal regulations [3]. This, along with the generally unpredictable market described in the section above, has led to an increase in interest of the public, and thus more information, studies and predictions have been necessary.

Similar works have previously been done on the general housing market in Sweden, but with different data and factors in mind. It is still important to mention that house pricing is very common to analyse through linear regression – the general consensus seem to be that it is, in fact, possible to obtain a successful linear model within a reasonable confidence interval for predicting the prices.
1.3. Objective and aim
The objective of this project is to provide a tool for possible prediction of condominium pricing in Stockholm. Several previous studies have been carried out regarding apartment pricing and its correlation to the condominium characteristics. This study will combine the discoveries and conclusions made in previous studies, with new research regarding the internal factors that can have an impact on the apartment pricing. The previous studies done in this field will be viewed as a groundwork and will prime the research aimed to be done in this project – which aims to further validate the previously formed models and conceivably add more factors to it, that can be useful when evaluating the condominium pricing in Stockholm.

The aim of the project – to enable a strategy for preliminary pricing of condominiums in consideration of possible affecting factors – will be conducted using statistically significant data to yield a correspondence between the former and the market selling price.

The objective of this is to provide a tool, both for the customer and the merchant, in the process of apartment trading. Researching the possible affecting factors of such an uncertain market can contribute to a more conscious construction business. Therefore the project can both be of relevance for the consumer as well as seller.

1.4. Scope
The data relevant to this project are the factors that might have an influence on the pricing, and their correlation to the pricing. In this study, the analyzed variables will vary between the characteristics of the apartment. The internal factors that will be analyzed are the following variables: floor, monthly fee, number of rooms, living area, district, selling date, and construction year. There will be a geographical delimitation as the study will only take apartments in Stockholm city into consideration; without a geographical limitation, the study would become too extensive. Delimitations are made accordingly to Stockholms stad, and the districts in Stockholm municipality that are said to belong to Stockholm city together amount to 8 different postal codes, which are: 111, 112, 113, 114, 115, 116, 117, 118. The districts which compose these postal codes and that will be the observation districts for this study can be seen in the table
Although the districts composed by the postal code 111 were desired to include in this study, unfortunately the data specifically from this area was not available, hence this area was excluded even though it is considered to be part of Stockholm city. Furthermore, this study will strictly only study the pricing of condominiums, i.e. apartment units that are outright owned and independently sellable and not leased [6]. One thing to take into consideration is that the reported selling price for the condominiums (the response variable) is based on the estate-agent’s last bidding, and occasionally the seller will remove the ad for the condominium prior to registering the last valid bid – meaning – sometimes the registered selling price might not be accurate and the condominium might have been sold for a higher price. Although this is something to keep in mind, it is said to happen very rarely and these rare cases can not be detected in the data set. Hence no procedural change has been made referring to this minor inconvenience in the data set [7]. This study only handles data regarding the condominiums in the areas mentioned below, during the time periods of January 2014 to April 2019. The volatility of the time value of money has not been taken into account for the time period.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>112</td>
<td>Kungsholmen, Essinge Islands</td>
</tr>
<tr>
<td>113</td>
<td>Vasastaden, Hagaparken</td>
</tr>
<tr>
<td>114</td>
<td>Östermalm, North Djurgården</td>
</tr>
<tr>
<td>115</td>
<td>Gärdet, Djurgården, Hjorthagen</td>
</tr>
<tr>
<td>116</td>
<td>Södermalm, Danviksklippan</td>
</tr>
<tr>
<td>117</td>
<td>West Södermalm, Reimersholme, Långholmen, Liljeholmen, Gröndal</td>
</tr>
<tr>
<td>118</td>
<td>Central Södermalm, Årsta Islets</td>
</tr>
</tbody>
</table>

Table 1.1: The postal codes included in this project, and their respective districts

### 1.5. Tools

Throughout the duration of the project, different programs were used. **Python** was used to obtain and convert the data set from Booli’s API. **Microsoft Excel** was used to store the data in a format to enable mathematical analysis. For the mathematical analysis and the method in general, **R** was used.
1.6. Problem formulation
The problem statement that will be answered in this thesis is:
Which factors are relevant to valuing condominiums, and to what extent?

2. Theoretical framework and modelling theory
2.1 An introduction to multiple linear regression
The multiple linear regression approach is based on modelling the relationship between a response variable (the dependent variable) \( y \) and multiple regressor variables \( x_1, x_2, \ldots, x_n \) by fitting a linear equation to observed data [8]; meaning that a regression model involving more than one regressor variable is called a multiple regression model. In general, the response \( y \) may be related to \( k \) regressors, covariates, or predictor variables. The model shown below is called a multiple linear regression model with \( k \) regressors. The parameters \( \beta_j, j = 0, 1, \ldots, k \) are called the regression coefficients. This model describes a hyperplane in the \( k \)-dimensional space of the regressor variables \( x_i \). The parameter \( \beta_j \) represents the expected change in the response \( y \) per unit change in \( x_j \) when all of the remaining regressor variables \( x_i (i \neq j) \) are held constant. For this reason the parameters \( \beta_j j = 1, 2, \ldots, k \), are often called partial regression coefficients. The error terms are normally distributed and denoted as \( \varepsilon_i \).

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i = \\
= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \ldots, n
\]

(eq. 2.1)

When working with multiple regression models it is often more convenient to use matrix notation; this allows a very compact display of the model, data, and results. In matrix notation the model given by Eq. (2.1) is

\[
Y = X\beta + \varepsilon
\]

(eq. 2.2)

Where \( Y \) is a \( n \times 1 \) vector containing the \( n \) response variables:

\[
Y = \begin{bmatrix}
y_1 \\
\vdots \\
y_n
\end{bmatrix}
\]

(eq. 2.3)
X is a $n \times (k + 1)$ matrix containing the $x_{nk}$ regressors:

$$
X = \begin{bmatrix}
1 & x_{11} & \cdots & x_{1k} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & \cdots & x_{nk}
\end{bmatrix}
$$

(eq. 2.4)

$\beta$ is a $(k + 1) \times 1$ vector containing the $k+1$ regression coefficients:

$$
\beta = \begin{bmatrix}
\beta_0 \\
\vdots \\
\beta_k
\end{bmatrix}
$$

(eq. 2.5)

$\varepsilon$ an $n \times 1$ vector containing the $n$ error terms:

$$
\varepsilon = \begin{bmatrix}
\varepsilon_1 \\
\vdots \\
\varepsilon_n
\end{bmatrix}
$$

(eq. 2.6)

### 2.1.1. Necessary assumptions

When using a multiple linear regression model, there are several assumptions that are needed to be taken into consideration. The used data needs to meet these assumptions in order for the analysis to be reliable and valid [9].

1. The first assumption states that the relationship between the response variable $y$ and the regressors $x$ is linear or near-linear. A typical violation of this assumption is the use of the incorrect regressors in the model, which will thus give a nonlinear relationship.

2. The second assumption states that the error term has a mean of zero.

3. The third assumption states that the error terms all have the same variance. A violation to this assumption would be the detection of heteroscedasticity, which means that the error terms do not have the same variance. This concept, and homoscedasticity, is explained in the section below.

4. The fourth assumption states that the errors are uncorrelated.

5. The fifth assumption states that the errors are normally distributed.
2.1.2. Homoscedasticity and heteroscedasticity

Homoscedasticity and heteroscedasticity refer, respectively, to whether the variances of the predictions determined by regression remain constant or if they differ [10]. These two problems that can occur when performing linear regression on a set of data are explained below.

2.1.2.1. Homoscedasticity diagnostics

The third assumption states that the error terms are uncorrelated and all have the same variance, this is called homoscedasticity and can be mathematically expressed as \( \text{Var}(\epsilon_i \mid x_i) = \sigma^2 \), where \( \epsilon_i \) is the error term, \( x_i \) is the regressor, and \( \sigma^2 \) the variance. A test of homoscedasticity of the error terms is desirable since it determines whether a regression model’s ability to predict a response variable is consistent across all values of that response variable. [11]

2.1.2.2. Heteroscedasticity diagnostics

Heteroscedasticity, the opposite of the concept explained above, can be described in mathematical terms as \( \text{Var}(\epsilon_i \mid x_i) = f(x_i) \), it is where the error term can be formulated as a function of \( x_i \). When heteroscedasticity occurs, the variance may often depend on the values of one or more of the explanatory variables, or on additional relevant quantities such as time or spatial ordering. [12]

2.1.2.3 Detecting heteroscedasticity

There are multiple ways of testing constant error variance, and thus also multiple ways of identifying heteroscedasticity within a model fit. Identifying heteroscedasticity can be done either statistically through different measures, in which a few will be explained later, or graphically, as it usually is sufficient enough to visually interpret plots [13].

The easiest way to detect heteroscedasticity is through a graphical measure, most commonly by interpreting the plot of residuals against the fitted values. For a model fit to be homoscedastic, there should be an random, equal distribution of residual points throughout the range of the fitted value x-axis [14].
For an added layer of justification, the statistical measures can be done. There are a multitude of different methods — the F-test, Modified Levene test and Bartlett’s test, amongst others. One of the most common tests of non-constant error variance is the Breusch-Pagan test, which is also what will be used in this study.

The Breusch-Pagan test assumes that the error terms are normally distributed, and that the null hypothesis is homoscedastic. The test is achieved by regressing the squared residual on the predictor, which results in a sum of squares, denoted SSR*, that provides a measure of dependency of the error term on the predictor [15].

The test statistic is mathematically calculated through the formula below,
\[ X^2 = \frac{SSR^*}{(SSE / n)^2} \] (eq. 2.7)
in which SSE is the sum of squared errors of prediction. The p-value for for this test is found using a $X^2$-distribution with one degree of freedom. For a model with a confidence interval of 95%, it can be concluded that a p-value of more than 0.05 in significance level is enough to reject the null hypothesis, that the model has non-constant error variance, and therefore that heteroscedasticity is present.

**2.1.3. Ordinary least squares estimation**

Presupposed that the assumptions mentioned above are applicable to the way in which the data are generated, the ordinary least-squares (OLS) method will be considered the optimal estimator of the unknown parameters $\beta$. The OLS estimator of $\beta$ is expressed as $\hat{\beta}$ and is achieved by minimizing the sum of squared errors. This is done mathematically by setting the derivative of the sum of the squared errors with respect to $\hat{\beta}$ equal to zero, which yields the least squares estimator of $\beta$ as follows:
\[ \hat{\beta} = (X'X)^{-1} X' y \] (eq. 2.8)
provided that the inverse matrix $(X'X)^{-1}$ exists. This matrix will always exist if the regressors are linearly independent, i.e. if no column of the $X$ matrix is a linear combination of the other columns. The least-squares estimator $\hat{\beta}$ is an unbiased estimator of $\beta$ under the multiple linear regression assumptions mentioned earlier, hence $E(\hat{\beta}) = \beta$. The variance property of $\hat{\beta}$ is as follows:
\[ \text{Cov}(\hat{\beta}) = \text{Var}(\hat{\beta}) = \sigma^2 \left( X'X \right)^{-1} \] (eq. 2.9)
2.1.4. Dummy variables

In regression analysis, the variables employed are often quantitative variables, meaning that the variables have well-defined scales of measurements. In this case, variables such as the monthly fee, floor, and living area are quantitative variables. In some situations it is necessary to use covariates measuring quality or category – attributes rather than quantities. In general, a qualitative variable has no natural scale of measurement. An example of a qualitative variable in this case is what time of year the condominium was sold – it is a qualitative characteristic but it has no intrinsic meaning of its own. To be able to account for the effect that the qualitative variables have on the response variable, the assignment of a set of levels is needed. This is done through the use of dummy variables.

A dummy is thus an artificial variable constructed to take either of the values 0 or 1, whenever one phenomenon or the other occurs. Mathematically we let the dummy variable take the value 1 if the respective state is occurring, and 0 if not. For example, regarding the dummy variables specifying during what season the condominium was sold, the states of the dummy specifying if it was sold during the summer or not will take two states depending on the observation. More specifically:

\[ x_i = \begin{cases} 
1 & \text{if it was sold during summer} \\
0 & \text{if not} 
\end{cases} \]

There are usually several different qualitative variables that must be incorporated into the model, often resulting in the need of more than one indicator – which is the case in this thesis. Continuing on the example of the dummy variables representing during which season the condo is sold, it is known that there are four seasons in a year – spring, summer, autumn and winter. Generally, if there are \( k \) related qualitative variables (\( k = 4 \) for the number of seasons in a year), only \( k-1 \) variables are assigned dummies. That is because having \( k \) dummies to represent the variables is redundant as it carries no new information, and will thus create a severe multicollinearity problem for the regression analysis (see section 2.3 for more about multicollinearity) [16]. If the seasons spring, summer and autumn are each assigned dummies and they all take on the value 0 – i.e. they were not sold during any of those seasons – then the only plausible explanation is that the condo was sold during the winter. The winter variable is in this case the benchmark for the other variables. Though, it is important to note that this only works for related qualitative variables that are mutually exclusive (not overlapping) and exhaustive (no other levels exist for this variable). [17]
2.2 Diagnostics of outliers and influential observations

An outlier is an extreme observation, a point that is considerably different from the rest and majority of the observations. If an observations falls far from the line implied by the rest of the data, it can be assumed to be an outlier because of a “bad value” that has resulted from a data recording or some other error. However, this data point can also be a valid value and may be a highly useful piece of evidence concerning the model generation and data analysis. Depending on their location in the system, outliers can have moderate to severe effects on the regression model, which can be considered positive or negative depending on the earlier discussion about whether or not it is a bad/valid value. Outliers should be carefully investigated to see if a reason for their unusual behaviour can be found, the “bad” value case occurs usually as a result of unusual but explainable events such as faulty measurements and analysis, and incorrect recording. If a “bad” value is the case, this observation should be corrected if possible and deleted from the data set if not. However, if the outlier is an unusual but perfectly valid observation deleting it to improve the fit of the equation would alter the predictive ability of the model and give a false sense of precision and accuracy. Additionally, outliers can sometimes be of great importance to include in the model compared to the rest of the data since they may control many key model properties. They may also point out inadequacies in the model, and expose the reality which may be that the linear relationship maybe is not as prominent as anticipated. [15]

2.2.1 DFFITS

The DFFITS is a measure of influence, formally defined as the studentized DFFITS, and calculated though:

\[
\text{DFFITS} = \frac{\hat{y}_i - \hat{y}_{(i)}}{s_i \sqrt{h_i}} \quad (\text{eq. 2.10})
\]

where the latter (DFFITS) is the change in the predicted value of a data point \(i\), obtained when that point \(i\) is left out of the regression, and the former (the studentization) is the division by the estimated standard deviation of the fit at that point [18]. A large value indicates that the observation is very influential in its neighborhood of the \(X\) space, and the data point should thus be considered for cutoff [19]. The size-adjusted cutoff is given by the absolute value of the formula:

\[
2 \sqrt{\frac{p}{n}} \quad (\text{eq. 2.11})
\]
where p is the number of parameters in the model, and n is the number of observations used to fit the model [20].

2.2.2 Covariance ratio

The Covariance ratio (covratio) is yet another approach to measure influence. The covariance ratio measures the impact of each observation on the variances and standard errors of the regression coefficients, and their covariances [21]. The covariance ratio is given by:

\[
\text{COVRATIO} = \left( \frac{S_0}{MS_{\text{Res}}} \right)^p \left( \frac{1}{1 - h_i} \right)
\]  
(eq. 2.12)

A high leverage point will generally make the covratio large. It is suggested that with a covratio larger than the value calculated though formula below, should be considered for cutoff [22].

\[
\left| \text{COVRATIO} - 1 \right| > \frac{3p}{n}
\]  
(eq. 2.13)

2.3. Model transformation

Model transformations might be needed when a data set is implied to be nonlinear, have non-constant variance and/or non-normal. Transforming a model actually means a transformation of a variable, and involves using mathematical operations to change a models’ measure scale. [23]

A transformation of a variable can be done on either one or multiple regressor variables \(x_i\), or on the response variable \(y\). The transformation method used in this study is called the Box-Cox transformation, as it is one of the most commonly used transformation methods, and also intuitive and simple to use.

The Box-Cox transformation is a transformation of the response variable \(y\).

2.3.1 Box-Cox transformation

If the desire is to transform \(y\) to correct non-normality or non-constant variance, one can use power transformation of \(y\) wherein \(y\) is raised to a certain power \(\lambda\). Both the parameters of the regression model and the exponent \(\lambda\) can be estimated simultaneously by the maximum likelihood method. This can be done by maximizing the equation shown below.
An approximate $100(1-\alpha)$ percent confidence interval for $\lambda$ are the $\lambda$ values that satisfy:

$$L(\hat{\lambda}) - L(\bar{\lambda}) \leq \frac{1}{2} \chi^2_{a,1}$$  \hspace{1cm} (eq. 2.15)

The confidence interval is constructed by plotting $L(\lambda)$ versus $\lambda$ and placing a horizontal line at height:

$$L(\hat{\lambda}) - \frac{1}{2} \chi^2_{a,1}$$  \hspace{1cm} (eq. 2.16)

A 95% confidence interval is then acquired by setting $\alpha = 0.05$

### 2.4 Multicollinearity

The purpose of regression analysis is to estimate the parameters of a dependency, not an interdependency; with other words – a relationship. [24] If there is no linear relationship between the regressors, they are said to be orthogonal. However, in most applications of regression models, the regression coefficients are not orthogonal. A significant problem that may dramatically impact the practicality of a regression model is the concept of multicollinearity, which implies that there are near-linear dependencies among the regressors. If there is strong multicollinearity between two regressors, the least-squares estimators of these regression coefficients will have large variances and covariances. Another consequence of multicollinearity is that it tends to produce least-squares estimates $\hat{\beta}_j$ that are too large in absolute value, algebraically this implies that the vector $\hat{\beta}$ is generally longer than the vector $\beta$. An important aspect to point out is that while the method of least squares will generally produce poor estimates of the individual model parameters when multicollinearity is present, this does not necessarily mean that the fitted model is a poor predictor.
2.4.1. Detecting multicollinearity
Several techniques can be employed to detect multicollinearity, the most common one being the VIF method, as it is easy to compute, intuitive, easy to understand, and useful when investigating the specific nature of the phenomenon in question. Below are a few of the possible methods explained.

2.4.1.1. Variance inflation factors
The main diagonal elements of the \((X^T X)^{-1}\) matrix are often called variance inflation factors (VIFs), these elements are very useful when the goal is to detect multicollinearity. In general, the VIF for the j:th regression coefficient can be mathematically written as:

\[
VIF = \frac{1}{1 - R_{j}^2} \quad (eq. \ 2.17)
\]

Where \(R^2\) is the value represented by the \(R^2\) obtained for doing a regression with each and every regression variable \(x_j\) with the rest of the variables as independent ones, mathematically regressed as:

\[
x_j = c_0 + x_{j+1} + \ldots + x_n + \epsilon \quad (eq. \ 2.18)
\]

One or more large VIFs indicate multicollinearity, a general rule is that if VIF > 5 the covariate should be examined further, and if any VIF > 10 this is a clear indication of multicollinearity among the regressors.

2.4.1.2. Linear dependence
Another way of explaining multicollinearity is that the values of one regressor variable in a regression model can be linearly predicted through another or more regressors [25]. A simple way of detecting multicollinearity can thus be done by constructing a scatter plot of the suspicious variables. By plotting the observations of the covariates against each other, possible linear dependency can be detected if the data points in the plot are scattered around a straight line.

2.4.1.3. Remedies to multicollinearity
While there are recommended ways to remedying multicollinearity, there is no definite solution to the problem. The easiest way is to remove the correlated predictors from the model.
If two factors have a high enough VIF-value, removing one of the variables from the model could solve the problem without significantly affecting the $R^2$-value of the model, since the regressor does not supply any new information. [26]

### 2.5. Variable selection methods

In most cases, including in this thesis, there is a rather larger group of possible candidate regressors, of which only a few are likely to be important regarding the response variable and the prediction aimed to achieve with the applied model. Finding an appropriate subset of regression variables for the applied model is usually called the variable selection problem. Good variable selection is very important if multicollinearity is present. The most common corrective treatment for multicollinearity is variable selection, this is intuitive since multicollinearity exists between the regression coefficients and if these are selected accordingly, the problem of multicollinearity can be eliminated. Although variable selection does not guarantee the elimination of this problem, since there are cases where two or more regressors are highly related, yet they strongly belong in the model. The variable selection method helps justify the presence of the highly correlated regression coefficients. Generating a regression model that only includes some of the regressors that were initially in the model involves two conflicting objectives:

1. It is preferable to have a model that includes as many regressors as possible so that the information content in these factors can influence the response variable $y$.
2. It is preferable to have a model that includes as few regressors as possible because the variance of the response variable $y$ increases as the number of regressors increase. Furthermore, the more regressors there are in a model, the greater the costs of data collection and model maintenance.

During the process of finding a well-fitting model one has to compromise between the above mentioned objectives, this compromise is called selecting the “best” regression equation. The word *best* is quoted since in this case there is no unique definition of the *best*. In most cases, there is usually not a single best equation but rather several equally good ones. Since variable selection algorithms are heavily computer dependent, there sometimes can be a temptation to place too much reliance and attention on the results of these procedures. Such temptation should be avoided, and experience, judgement, intuitiveness and even subjective considerations should be used as methods to explore the structure of the data and to draw conclusions. [27]
2.5.1. All possible regression

To employ the best subset of the total regressor variables, the method of all possible regression has been chosen. This procedure tests all possible subsets of the set of independent variables; more intuitively – this procedure fits all possible models based on the independent variables and after fitting all of the models, the best models with one independent variable, then two independent variables, then three (and so on), are displayed. Assuming the intercept term $\beta_0$ (which is constant) is included in all equations, with K candidate regressors there are $2^K$ separate equations in total to be examined. [28] All-possible-regression procedures will efficiently process up to about 30 candidate regressors with computing times that are comparable to other methods for variable selection [29]. The least squares regression method is used to estimate the coefficients of the variables that are kept. There are a number of different criteria to rank the best subset models that have been generated with the all-possible-regression model [30], the following criteria will be used in this thesis:

1. Adjusted $R^2$
2. Bayesian information criterion (BIC)

Below are descriptions of each and every criterion.

2.5.2. Adjusted $R^2$

A way to assess the overall adequacy of the model, denoted $R^2_{\text{Adj},p}$ will only increase on adding a variable to the model if the addition of the variable reduces the residual mean square. This statistic does not necessarily increase as additional regressors are introduced into the model. In fact, if $s$ regressors are added to the model, $R^2_{\text{Adj},p+s}$ will exceed $R^2_{\text{Adj},p}$ if and only if the partial F statistic for testing the significance of the $s$ additional regressors exceeds 1. Hence, one criterion for selection of an optimum subset model is to choose the model that has a maximum $R^2_{\text{Adj},p}$. Defined as:

$$R^2_{\text{Adj},p} = 1 - \left( \frac{n-1}{n-p}\right)(1 - R^2_p)$$

(eq. 2.19)

2.5.3. Bayesian information criterion

Akaike proposed an information criterion, AIC, based on maximizing the expected entropy of the model, which is a measure of the expected information. The Bayesian information criterion (BIC) places a greater penalty on adding regressors as the sample size increases, than the AIC.
The BIC is a criterion for model selection among a finite set of models, it is partly based on the likelihood function and it closely related to the Akaike information criterion (AIC). [31] The BIC is given by:

\[
\text{BIC} = \frac{1}{n} \left( \text{RSS} + \log(n)\hat{\sigma}^2 \right) \quad \text{(eq. 2.20)}
\]

2.6. Model validation
Since regression models are frequently used for prediction, estimation, data description and control, not only by the model developer but mainly other individuals, the generated model in studies such as this one should be assessed of their validity before being released to the public. Model validation is directed toward determining if the model will function successfully in its intended operating environment, in this case predicting the selling price of condominiums in Stockholm city. [32]

A good method of model validation is the collection of new data which can be compared with the predictions of the model. The validity of the mathematical and physical assumptions used in developing a model and estimating the coefficients is less open to question if a model gives accurate predictions of new data. [33]

2.6.1. Hypothesis testing
Hypothesis testing is a method used in statistics to make decisions using experimental data. The key terms and concepts, including the four elemental steps for the testing are:

1. Formulating the null hypothesis \( H_0 \) and the alternative hypothesis \( H_1 \)
   \( H_0 \) is a statistical hypothesis that assumes that the observation is due to a chance factor, meaning it is the result of pure chance. \( H_1 \) shows that the observations are the result of a real effect and of chance variation
2. Identify a test statistic that can be used to assess the truth of \( H_0 \)
3. Compute the p-value – the probability that a test statistic at least as significant as the one observed would be obtained assuming that \( H_0 \) is true. The smaller the p-value, the stronger the evidence against \( H_0 \)
4. Compare the p-value to an acceptable significance value \( \alpha \) (also called an alpha value). If \( p \leq \alpha \), the observed effect is statistically significant and the null hypothesis is ruled out. Hence the alternative hypothesis is valid [34]
2.6.2. p-value
The p-value is the probability that a variate would assume a value greater than or equal to the observed value strictly by chance. When the p-value is used in statistics for hypothesis evaluation, the null hypothesis $H_0$ is set to be that the corresponding regression coefficient $\beta_i$ is equal to zero. The alternative hypothesis is then set to be that the corresponding coefficient is not equal to zero, respectively. The p-value is a number between 0 and 1 and interpreted in the following way when the chosen significance level $\alpha$ is 5%:

- A small p-value (typically $< 0.05$) indicates strong evidence against the null hypothesis $H_0$.
- A large p-value (typically $> 0.05$) indicates weak evidence against the null hypothesis, so the null hypothesis is not rejected.
- P-values very close to the turning point (0.05) are considered to be marginal and could be considered to go either way, these are left open for discussion [35].

The p-value in itself does not support reasoning about the probabilities of hypotheses, but is a tool for deciding whether to reject the null hypothesis $H_0$ or not. [36]

2.6.3. Out of sample validation
A good way to test the assumptions of a model and to realistically compare its forecasting performance against other models is to perform out of sample validation, which means to withhold some of the sample data from the model identification and estimation process, to then use the model to make predictions for the hold-out data in order to see how accurate the prediction is in comparison to the actual data. [37]

3. Method
The main mathematical method used in this thesis is multiple linear regression analysis, described in section 2, with the intention to generate a model aimed to predict the pricing of condominiums in Stockholm city. In order to successfully generate a model, data is first collected and preprocessed for the experimental setup. Then, the model for prediction can be brought forth and honed with the help of relevant statistical tools. Lastly, the validity of the model is analyzed.
3.1. Initial data set and preprocessing

To be able to generate a predictive model a large amount of observations were collected, this data was provided by Booli Search Technologies AB through their website https://www.booli.se/p/api and their database available for research purposes. The data was reclaimed from their application programming interface (API) through the login-details provided for the use of this specific thesis, by e-mail communication with Mikaela Lagercrantz, developer at Booli.

By using Booli’s API, sales from January 2014 until April 2019 were collected. These were converted from their database to a file in Microsoft Excel with a code written in Python. A total of 10 472 observations were collected from the postal codes comprising Stockholm city. As mentioned before, the postal codes were 112, 113, 114, 115, 116, 117, 118. The 10 472 observations of sales contained data regarding:

➢ Selling price
➢ Construction year
➢ Floor
➢ Number of rooms
➢ Monthly fee
➢ Living area in m²
➢ Distance to the ocean
➢ Coordinates: longitude/latitude
➢ District
➢ Address
➢ Estate agency
➢ Estate agent
➢ Additional area (such as balcony, loft, storage-room)
➢ Listing price
➢ Publishing date
➢ Selling date

Upon reviewing and analyzing the available data, the observations were narrowed down into a group of fewer regressor variables. Starting from the top of the list, here are the included and excluded variables and their explanatory statements, respectively: selling price was included since this is the response variable and has to be included for the analysis of the linear relationship, construction year was
included as a regressor since different time periods represent different styles of construction and might thus have an effect on the dependent variable, floor, number of rooms, monthly fee and living area were all included as regressor variables since they are suspected to have an impact on the response variable. Distance to the ocean was excluded since the variable district (of the city) was included instead, as this factor has a higher possibility of showing significantly larger impact on the response variable. Coordinates and address were removed since these qualities are represented by the variable district, already included in the initial model.

After careful contemplation, the variables estate agency and estate agent were removed. It is widely known in some communities and areas of cities that certain estate agents and brokers announce particularly high/low starting bids. Active members of the market including customers and even competitors, know of this phenomenon and although it is a surprising factor that affects the value of the property, there seems to be a correlation between the selling price being unusually high/low and the estate agent that is of question. We considered this phenomenon to be of much greater value and importance than what this project would have assessed it, a whole study on solely these factors could be done to analyze the situation to give it justice. Hence the variables estate agent and estate agency were rightfully removed from the data set.

The additional area was removed since this was a characteristic that was available for such few data points, that these observations would not have carried any significance for the mathematical analysis and the correlation between this specific regressor and the response variable. Lastly, the publishing date was removed since the selling date was taken into consideration instead, represented by dummy variables.

<table>
<thead>
<tr>
<th>Regressor variable</th>
<th>Notation</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>Floor</td>
<td>Floor number</td>
<td>On what floor of the building the condominium is situated on. Ranges between 0 and 25 floors.</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>Fee</td>
<td>SEK</td>
<td>The monthly fee paid as part of the maintenance costs of the building and to the corresponding association for being part of the community of condominiums in that area. The monthly fee of zero SEK is usually because the association has paid off their loans, or because it earns a significant enough amount of commercial revenue to cover the residents’ fees [38].</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>Rooms</td>
<td>Number of rooms</td>
<td>How many rooms there are in the condominium. This variable can in some cases have a value with a decimal (,5) and means that case that there is an alcove of the room that counts as “half” of a room. Ranges between 1 and 9 rooms.</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>Area</td>
<td>( m^2 )</td>
<td>How many ( m^2 ) the condominium is. Ranges between 12 and 333 ( m^2 ).</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>Spring</td>
<td>Dummy</td>
<td>Where spring is defined as the period between March and May.</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>Summer</td>
<td>Dummy</td>
<td>Where summer is defined as the period between June and August.</td>
</tr>
<tr>
<td>( x_5 )</td>
<td>Autumn</td>
<td>Dummy</td>
<td>Where autumn is defined as the period between September and November.</td>
</tr>
<tr>
<td>Winter</td>
<td>Benchmark</td>
<td></td>
<td>Where winter is defined as the period between September and November.</td>
</tr>
<tr>
<td>( x_6 )</td>
<td>Con1</td>
<td>Dummy</td>
<td>Condominium buildings with construction year 1919 and older, where the oldest is constructed year 1650.</td>
</tr>
<tr>
<td>( x_7 )</td>
<td>Con2</td>
<td>Dummy</td>
<td>Condominiums buildings with construction year 1920 - 1949.</td>
</tr>
<tr>
<td>( x_8 )</td>
<td>Con3</td>
<td>Dummy</td>
<td>Condominiums buildings with construction year 1950 - 1999.</td>
</tr>
<tr>
<td>Con4</td>
<td>Benchmark</td>
<td></td>
<td>Condominiums in buildings with construction year 2000 and forward.</td>
</tr>
<tr>
<td>( x_9 )</td>
<td>Dis1</td>
<td>Dummy</td>
<td>District with postal code 113 – Vasastan, Hagaparken</td>
</tr>
<tr>
<td>( x_{10} )</td>
<td>Dis2</td>
<td>Dummy</td>
<td>District with postal code 114 – Östermalm, North Djurgården</td>
</tr>
<tr>
<td>( x_{11} )</td>
<td>Dis3</td>
<td>Dummy</td>
<td>District with postal code 116 – Södermalm, Danviksklippan</td>
</tr>
<tr>
<td>( x_{12} )</td>
<td>Dis4</td>
<td>Dummy</td>
<td>District with postal code 118 – Central Södermalm, Årsta Islets</td>
</tr>
<tr>
<td>( x_{13} )</td>
<td>Dis5</td>
<td>Dummy</td>
<td>District with postal code 115 – Gärdet, Djurgården, Hjorthagen</td>
</tr>
<tr>
<td>( x_{14} )</td>
<td>Dis6</td>
<td>Dummy</td>
<td>District with postal code 117 – West Södermalm, Reimersholme, Långholmen, Liljeholmen, Gröndal</td>
</tr>
<tr>
<td>Dis7</td>
<td>Benchmark</td>
<td></td>
<td>District with postal code 112 – Kungsholmen, Essinge Islands</td>
</tr>
</tbody>
</table>

Table 3.3: The regressor variables included in the initial model
Additionally, not only a few of the regressors, but a large portion of the observations were removed too. This because they did not possess all the characteristics that are listed above, thus these observations are considered to be inadequate data regarding their contribution to the model. Elaborating further on this, there was a total of 2317 data points with missing data regarding *monthly fee*, *living area*, *number of rooms*, or *construction year*. Removing these observations with the earlier explained reasoning, the data set was reduced to a total of 8155 observations. Additionally, removing the outliers resulted in a final set of data consisting of 8040.

A quick overview of the plots below in table 3.1, with the *initial data set* (the set of 10472 observations) shows sporadic observation points, and parts of the data that clearly do not follow the general trend. The five plots shown below give an overlook of the general state of the data set, obtained with the *lm*-function in R. After removing the observations that were mentioned and explained earlier, this general overlook through the *lm*-function gives a rather more homogenous and expected distribution of the data points in the plots (see table 3.2). Especially one problem area that diminishes after removing the observations with missing characteristics, is the clear line that can be seen in the “residuals vs. fitted” plot, and the “scale-location” plot. Once again, this is just a general way of depicting the convenience of removing these observations and should not be viewed as a thorough analysis. Variable selection and different models for analyzing the data can be read further on in the project.

![Plots from table 3.1](image)

Table 3.1: General state of the data, with the set of 10472 observations
Table 3.2: General state of the data, with the set of 8040 observations

Below is a histogram, representing the distribution of the numerical data regarding the buildings’ construction year. Historical data shows that different periods of time represent different characteristics and architectural styles – everything from used materials, building foundation, ceiling height, to window and door framings. The grouping of time periods was made to fit the ratio between frequency and interval size, meaning that the group intervals are smaller in size if the number of buildings built during a certain period are certainly large, and the period is larger in size if there were particularly few buildings built during that era. The blue vertical lines represent the points in which the intervals are divided into separate ones. The x-axis only goes back to 1850, leading to some missing values from earlier years – this data would neither affect the histogram nor the grouping since the number of constructed buildings is particularly low between the starting year (1650) and the earliest shown construction year in the plot below, being 1850.
Graph 3.1: Histogram showing the frequency of constructed buildings between 1850-2019

3.2. Variable selection
It is desired for the significance level of the final model to be above 95%, i.e. $\alpha = 0.05$, which is why the optimal model should not include those variables or dummies with a p-value above 0.05. Four variables were removed based on the p-test, these consisted of: spring, autumn, con3, and rooms. As mentioned earlier in the report, the method of all possible regression and the BIC were also used as methods for excluding the appropriate variables and selecting the entitled ones. These last two methods mentioned alone led to the deletion of only one variable, which was dis3. The excluded variables are presented in a table below. Finally, this explained method and the results are discussed further in section 5, the analysis and discussion.

<table>
<thead>
<tr>
<th>Regressor variable</th>
<th>Notation</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_3$</td>
<td>Rooms</td>
<td>No. of rooms</td>
<td>How many rooms there are in the condominium.</td>
</tr>
<tr>
<td>$x_5$</td>
<td>Spring</td>
<td>Dummy</td>
<td>Where spring is defined as the period between March and May.</td>
</tr>
<tr>
<td>$x_7$</td>
<td>Autumn</td>
<td>Dummy</td>
<td>Where autumn is defined as the period between September and November.</td>
</tr>
<tr>
<td>$x_9$</td>
<td>Con2</td>
<td>Dummy</td>
<td>Condominiums buildings with construction year 1920 - 1949.</td>
</tr>
<tr>
<td>$x_{13}$</td>
<td>Dis3</td>
<td>Dummy</td>
<td>Södermalm</td>
</tr>
</tbody>
</table>

Table 3.3: The excluded variables and their informations, respectively
3.3 Model validation
A total of 688 observations were obtained by randomizing observations from a data set that was not used to generate the model.

4. Results

4.1. The final model
The obtained final model of this study included the quantitative variables floor, area and fee, and the dummy variables con1, con2, con3, dis1, dis2, dis4, dis5, dis6. The regression equation given was:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_{10} x_{10} + \beta_{11} x_{11} \\
+ \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{14} x_{14} + \beta_{15} x_{15} + \beta_{16} x_{16} \]  (eq. 4.1)

The above presented mathematical expression can be rewritten as a less formal one, by inserting the regression coefficients as numbers and the variables with their respective names. Below is the values given in R, and thereafter the less formal version of the final model, with the coefficients rounded off to integer numbers.

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>Floor</th>
<th>Fee</th>
<th>Area</th>
<th>Con1</th>
<th>Con3</th>
<th>Dis1</th>
<th>Dis2</th>
</tr>
</thead>
<tbody>
<tr>
<td>543948.7819</td>
<td>51539.8635</td>
<td>-248.4169</td>
<td>82115.7392</td>
<td>371466.2224</td>
<td>-510838.3662</td>
<td>306210.5984</td>
<td>775792.6610</td>
</tr>
<tr>
<td>Dis3</td>
<td>Dis4</td>
<td>Dis5</td>
<td>Dis6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64791.1783</td>
<td>-1422713.1557</td>
<td>220450.8264</td>
<td>-1028799.7539</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: The results of the final model in R

\[
selling \text{ price} = 543\,949 + 51\,531 \cdot \text{floor} \\
- 248 \cdot \text{fee} \\
+ 82\,116 \cdot \text{area} \\
+ 371\,466 \cdot \text{con1} \\
- 510\,838 \cdot \text{con3} \\
+ 306\,211 \cdot \text{dis1} \\
+ 775\,793 \cdot \text{dis2} \\
+ 64\,791 \cdot \text{dis3} \\
- 1\,422\,713 \cdot \text{dis4} \\
+ 220\,450 \cdot \text{dis5} \\
- 1\,028\,800 \cdot \text{dis6} \]  (eq. 4.2)
4.2. Comparisons with previous work

Comparing the obtained final model to previously analyzed and generated models can be of great value when evaluating the integrity of the model. One model, generated by Alexander Gustafsson and Sebastian Wogenius [39] as part of a thesis study reads as follows:

\[
\begin{align*}
\text{Price} &= 642060 \\
&= 60825 \times \text{Area} + 125570 \times \text{Balcony}_{3,0} + \text{ConstructionYear}_{0,1,2,3,4,5} + \text{District}_{0,1,2,3,4,5} + 155390 \times \text{Elevator}_{0,1} + 192340 \times \text{Fireplace}_{0,1} - 237800 \times \text{GroundFloor}_{0,1} - 210 \times \text{MonthlyFee} + 321420 \times \text{Penthouse}_{0,1} \\
&= 642060 + 60825 \times \text{Area} + 125570 \times \text{Balcony}_{3,0} + \text{ConstructionYear}_{0,1,2,3,4,5} + \text{District}_{0,1,2,3,4,5} + 155390 \times \text{Elevator}_{0,1} + 192340 \times \text{Fireplace}_{0,1} - 237800 \times \text{GroundFloor}_{0,1} - 210 \times \text{MonthlyFee} + 321420 \times \text{Penthouse}_{0,1} \\
\end{align*}
\]

Table 4.2: Linear model generated by another study

Firstly, it can be pointed out that the intercept in this model (642 060) is a number considerably close to the intercept generated in the model of this thesis (543 949). This is a desirable occurrence since the y-intercept ($\beta_0$) of an equation is a substantial part of the response variable, especially when the number is quite a large one – which it is in this case. Moreover, it is important to point out that the model presented in the mentioned study, has taken different variables into account; as for example, variables such as balcony, elevator, fireplace, and dummies for penthouse/ground floor have been incorporated into the model. These factors will have substantial impact on the overall look of the model and the coefficients in front. However, the coefficient for monthly fee (-210) and the coefficient for fee in the model generated in this study (-248) are exceptionally close to each other in value, as is the coefficient for area (60 825) and the one in this thesis (82 116). These are of course desirable comparisons and results, but it should be kept in mind that it could be accidental possibilities, synchronized by chance.
5. Analysis and discussion
This section of the thesis aims to analyze and discuss the presented theory, method, and results. Especially discussed is the calculations and theory behind the data transformation, the handling of outliers, the variable selection, and finally the model validation.

5.1. Box-Cox transformation
In the analysis of data it is often assumed that observations of the response variable $y_1, y_2, \ldots, y_i$ are independently normally distributed with constant variance and with expectations specified by a model linear with a set of parameters (see Necessary assumptions section 2.1.1.). Although, at times, a less restrictive assumption can be made that such a normal, homoscedastic, linear model is appropriate only after some suitable transformation has been applied to the response variable – i.e. by performing a power transformation on the response variable $y$ by raising it to a power $\lambda$. [40]

Plot 5.1: The standardized vs. studentized residuals plotted against the Box-Cox transformed fitted values, respectively

Plot 5.2: Q-Q plot of normality for the different model fits
5.2. Handling of outliers and influential observations

5.2.1. Residual analysis

A residual may be viewed as the deviation between the data and the fit, it is also a measure of the variability in the response variable not explained by the regression model. It is also convenient to think of the residuals as the realized or observed values of the model errors. Analysis of the residuals is an effective way to discover several types of model inadequacies, plotting the residuals is a very effective way to investigate how well the regression model fits the data and to check the assumptions (mentioned in section 2.1.1). Plotting different kinds of residuals can amount to a better analysis, four different residuals types are plotted and explained below – the “regular” residuals explained first and followed up by two kinds of scaled ones. [41]

5.2.1.1. Regular residuals

A plot of the regular residuals against the corresponding fitted values $\hat{y}_i$ is useful for detecting several common types of model inadequacies. The residuals are plotted versus the fitted values $\hat{y}_i$ and not the observed values $y_i$ since the $\varepsilon_i$ and the $y_i$ are uncorrelated while the $\varepsilon_i$ and the $y_i$ are usually correlated. The outward-opening funnel pattern that can be seen in this plot, implies that the variance is an increasing function of the response variable $y$, and is a pattern that is considered symptomatic of model deficiencies.

Plot 5.3: The regular residuals plotted against the fitted values
The usual approach for dealing with inequality of variance is to apply a suitable transformation to the regressors or the response variable. Plotting the residuals against the fitted values may also reveal one or more unusually large residuals. This residual plot would ideally show a more random pattern, which would aid in the aim to adapt a linear model to the set of data. Although, some points can be observed to have particularly large residuals in this plot and can potentially be outliers – they should be kept in mind throughout the project. If these points are recurrent during the plots of different kinds of residuals, transformations will be needed. [42] The regular residuals plotted in this case are non-scaled and are therefore not useful when comparison is needed, and since variances of the residuals may differ despite the fact that the variance of the true errors is equal – one would like to scale the residuals so that this is taken into consideration. This is mentioned below as the studentized residuals.

5.2.1.2. Studentized residuals

![Studentized Residuals against the Fitted Values](image)

Plot 5.4: The studentized residuals plotted against the fitted values

The studentized residuals have constant variance (\(\text{Var}(r_i) = 1\)) regardless of the location of \(x_i\) in the system if the model is correct. Hence the studentized residuals can be a more useful alternative when analyzing the residuals, since they may show violations of model assumptions at influential points that are more remote than the rest of the data. Since violations of the model assumptions are more likely at these remote points, violations may otherwise be harder to detect from inspection of the ordinary or other kinds of scaled residuals, since these residuals will usually be smaller.
5.2.1.3. Standardized residuals

![Plot 5.5: The standardized residuals plotted against fitted values](image)

The studentized residuals have constant variance, as mentioned before, regardless of the location of \( x_i \) if the model is correct, and in many situations the variance of the residuals stabilizes. In these cases there may be little difference between the standardized and studentized residuals – hence the standardized and studentized residuals often convey equivalent information. The standardized residuals have approximately unit variance and mean zero. Consequently, a large standardized residual could also be an indication of an outlier. If the standardized residual has a margin of ±2 it can be said that the frequency of the observed point is greater or less than the expected frequency. Residuals that are ±3 indicate that the point may be very unusual. This is motivated by the rule which specifies that 68, 95, and 99.7% of all values fall within one, two, or three standard deviations from the mean respectively if the observations are normally distributed. The vertical blue lines shown in the plot show the observations which exceed a standardized residual of ±3.

Although the theory regarding the standardized residuals states that the points with a residual absolute value larger than three are very unusual, the outliers that were decided to be deleted were the ones that had a residual absolute value considerably larger than three and were classified as outliers in the studentized residuals-plot. Meaning that, only the mutual outliers detected in the standardized- and the studentized plot that had a standardized residual absolute value larger than three, were removed from the data set. This because the theory implies that ±3 is a large standardized residual and observations with such values should be observed as unusual values, this led to the conclusion that one should be careful not to discard these points immediately.
Hence, observations with larger standardized residuals and the ones detected as outliers in the studentized plot were removed. By this method, 154 observations were discarded. This method, based on the theory of both the standardized and the studentized ones, can be said to yield safer treatment of outliers and influential points. Eventually, the data set was reduced to a total of 7076 observations.

The Adjusted $R^2$ and VIF factors were noted throughout the process of removing outliers. The Adjusted $R^2$ value rose overall, which was to be expected from removing far away outliers.

The VIF factors generally presented lower values than previous, with a few exceptions in which the differences were low enough to be deemed insignificant.

Table 5.1: Table of VIF-values for the initial model vs. the model without outliers

5.2.2. Covariance ratio and DFFITS

The methods considered for the diagnostic of outliers was primarily with the consideration of the cutoff values of the covariance ratio (covratio) and DFFITS. With the program R and its built in function infl, the covratio and DFFITS values could be calculated, in which all values for the individual data points acquired through the mentioned methods were seen as outliers and removed from the data if the value exceeded both the cutoff value of covratio and DFFITS.

The covratio cutoff value for the used data was 1.0063, rounded down to 1.00, and the cutoff value for DFFITS was 0.0913, rounded down to 0.09.

An example of the cut off outliers are data point 103, 265, 327, 452 and 458 presented in table 3.4 below.

Table 5.2: Part of the table of values acquired through R with the infl-function
Cook’s Distance is another commonly used measure of influence that was not exerted because the values acquired are indicators of the influence given in percentage. Because of the cast amount of data point used in this study, the values acquired through Cook’s distance were too low to fully analyze and utilize.

5.3. Variable selection

The method employed for variable selection was a combination of checking for multicollinearity, p-test, all possible regression, and the Bayesian information criterion. The excluded variables were discarded based on the combination of these methods, to make sure the deleted variables were identified by at least two different theoretical methods. Hence the joint discussion below, containing a combination of all theories.

5.3.1. Handling multicollinearity

Looking at the obtained VIF values from the regression model, it can be seen that there are two regressors – area and rooms – with values that exceed the recommended max value of 5.

<table>
<thead>
<tr>
<th>Floor</th>
<th>Rooms</th>
<th>Fee</th>
<th>Area</th>
<th>Spring</th>
<th>Summer</th>
<th>Autumn</th>
<th>Con1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0819</td>
<td>5.3601</td>
<td>3.1905</td>
<td>6.8474</td>
<td>1.4894</td>
<td>1.3764</td>
<td>1.5316</td>
<td>4.0866</td>
</tr>
<tr>
<td>Con2</td>
<td>Con3</td>
<td>Dis1</td>
<td>Dis2</td>
<td>Dis3</td>
<td>Dis4</td>
<td>Dis5</td>
<td>Dis6</td>
</tr>
<tr>
<td>4.7416</td>
<td>3.1864</td>
<td>1.9703</td>
<td>2.0068</td>
<td>2.2908</td>
<td>2.1921</td>
<td>1.8147</td>
<td>1.3523</td>
</tr>
</tbody>
</table>

Table 5.3: VIF-values for the regression fit without outliers

As mentioned in section 2.3.1, it is worth investigating the collinearity these two variables because of their VIF values. The first way this can be done is by examining the linear dependency between the two by conducting a scatter plot.

Plot 5.6: Scatter plot of the number of rooms against the area
By observing the plot shown above, one can notice a clear linear relationship between the two variables in general, despite a few irregularities. Adding to that, the VIF-values of the regressors showed a general decrease after the removal of rooms, as presented below.

<table>
<thead>
<tr>
<th>Floor</th>
<th>Fee</th>
<th>Area</th>
<th>Spring</th>
<th>Summer</th>
<th>Autumn</th>
<th>Con1</th>
<th>Con2</th>
<th>Con3</th>
<th>Dis1</th>
<th>Dis2</th>
<th>Dis3</th>
<th>Dis4</th>
<th>Dis5</th>
<th>Dis6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0814</td>
<td>3.1582</td>
<td>3.0073</td>
<td>1.4891</td>
<td>1.3754</td>
<td>1.5313</td>
<td>4.0830</td>
<td>4.7385</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Con3</td>
<td>Dis1</td>
<td>Dis2</td>
<td>Dis3</td>
<td>Dis4</td>
<td>Dis5</td>
<td>Dis6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1829</td>
<td>1.9696</td>
<td>2.0054</td>
<td>2.2879</td>
<td>2.1901</td>
<td>1.8146</td>
<td>1.3500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.4: VIF-values after removing the variable rooms

5.3.2. p-value, all possible regression, BIC

According to the p-values, the regressors to discard are spring, autumn and con2, as their p-values were well above the ideal value of 0.05. The variable spring showed a p-value of 0.23, autumn showed a value of 0.31, and con2 of 0.115, as presented in table 5.5 below. After removing the mentioned regressors, the summary of the new regression fit, displayed by table 3.6, showed no variables with a p-value high enough to consider for removal.

Table 5.5: Table summarizing the regression fit for the first iteration

<table>
<thead>
<tr>
<th>Coefficients: Estimate Std. Error t value Pr(&gt;</th>
<th>t</th>
<th>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>8.479e+01</td>
<td>2.151e-01</td>
</tr>
<tr>
<td>Floor</td>
<td>2.380e-01</td>
<td>1.838e-02</td>
</tr>
<tr>
<td>Fee</td>
<td>-7.334e-04</td>
<td>5.224e-05</td>
</tr>
<tr>
<td>Area</td>
<td>3.200e-01</td>
<td>2.425e-03</td>
</tr>
<tr>
<td>Spring</td>
<td>1.088e-01</td>
<td>9.038e-02</td>
</tr>
<tr>
<td>Summer</td>
<td>-2.777e-01</td>
<td>1.133e-01</td>
</tr>
<tr>
<td>Con1</td>
<td>-1.014e-01</td>
<td>1.005e-01</td>
</tr>
<tr>
<td>Con2</td>
<td>1.465e+00</td>
<td>1.675e-01</td>
</tr>
<tr>
<td>Con3</td>
<td>-2.389e-01</td>
<td>1.516e-01</td>
</tr>
<tr>
<td>Dis1</td>
<td>1.919e-01</td>
<td>1.376e-01</td>
</tr>
<tr>
<td>Dis2</td>
<td>3.117e+00</td>
<td>1.379e-01</td>
</tr>
<tr>
<td>Dis3</td>
<td>4.145e-01</td>
<td>1.233e-01</td>
</tr>
<tr>
<td>Dis4</td>
<td>-7.297e-01</td>
<td>1.386e-01</td>
</tr>
<tr>
<td>Dis5</td>
<td>9.351e-01</td>
<td>1.414e-01</td>
</tr>
<tr>
<td>Dis6</td>
<td>-6.686e+00</td>
<td>1.959e-01</td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 5.6: Table summarizing the regression fit for the second iteration
According to both stepwise regression methods, besides the previously mentioned covariates, the variable *summer* should be excluded from the model unless it were to include 12 or more variables.

Table 5.7: Results from the BIC with rows corresponding to the object and columns representing the number of parameters in the model and the BIC

<table>
<thead>
<tr>
<th>Selection Algorithm: backward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floor Fee Area Spring Summer Autumn Con1 Con2 Con3 Dis1 Dis2 Dis3 Dis4 Dis5 Dis6</td>
</tr>
<tr>
<td>1 (1)</td>
</tr>
<tr>
<td>2 (1)</td>
</tr>
<tr>
<td>3 (1)</td>
</tr>
<tr>
<td>4 (1)</td>
</tr>
<tr>
<td>5 (1)</td>
</tr>
<tr>
<td>6 (1)</td>
</tr>
<tr>
<td>7 (1)</td>
</tr>
<tr>
<td>8 (1)</td>
</tr>
<tr>
<td>9 (1)</td>
</tr>
<tr>
<td>10 (1)</td>
</tr>
<tr>
<td>11 (1)</td>
</tr>
<tr>
<td>12 (1)</td>
</tr>
<tr>
<td>13 (1)</td>
</tr>
<tr>
<td>14 (1)</td>
</tr>
<tr>
<td>15 (1)</td>
</tr>
</tbody>
</table>

Table 5.8: Identical to the previous table, but representing the results from the all possible regression method

<table>
<thead>
<tr>
<th>Selection Algorithm: exhaustive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floor Fee Area Spring Summer Autumn Con1 Con2 Con3 Dis1 Dis2 Dis3 Dis4 Dis5 Dis6</td>
</tr>
<tr>
<td>1 (1)</td>
</tr>
<tr>
<td>2 (1)</td>
</tr>
<tr>
<td>3 (1)</td>
</tr>
<tr>
<td>4 (1)</td>
</tr>
<tr>
<td>5 (1)</td>
</tr>
<tr>
<td>6 (1)</td>
</tr>
<tr>
<td>7 (1)</td>
</tr>
<tr>
<td>8 (1)</td>
</tr>
<tr>
<td>9 (1)</td>
</tr>
<tr>
<td>10 (1)</td>
</tr>
<tr>
<td>11 (1)</td>
</tr>
<tr>
<td>12 (1)</td>
</tr>
<tr>
<td>13 (1)</td>
</tr>
<tr>
<td>14 (1)</td>
</tr>
<tr>
<td>15 (1)</td>
</tr>
</tbody>
</table>
Observing plot 5.7, containing five different plots of different measures, it can be seen that the observable function flattens out in every plot right after passing the $x = 11$ point. These plots show that the predictive ability of the model slowly gets better with every regressor that is added to the regression, until 10 regressors are passed – the ability does not improve after that despite additional regressors. When deciding on removing one last regressor, the decision was to remove the `summer` variable, based on the results from the BIC and all possible regression-method. To further validate the deletion of this (last) regressor, the p-value can be pointed out to be considerably higher than all the other regressors in both iterations – although it is at an acceptable level (below 0.05), it is fair to say that it is larger than the values of the remaining regressors. (see table 5.5. and 5.6 for p-value iterations)
5.4. Model validation

5.4.1 Out of sample validation

Plot 5.8: The selling prices from the sample data set plotted against the fitted values obtained with the final model, where:

- represents the regression fit for the set of sample data
- represents 5% confidence-interval for the regression fit
- represents the ideal line on which the observations should lie on, if the final model would forecast with 100% certainty

By randomizing the set of samples to validate the model with, it is ensured that not any of the points that were used in developing the model are taken into consideration – this would be undesirable since the would yield a perfectly corresponding value as the original one – hence giving a false sense of predictive ability. However, an observation of plot 5.8 clearly shows a linear relationship between the original values (selling price from the sample data set) and the fitted values (selling price forecasted with the final model). The result could have shown an even better relationship with less distance between the data points and the line, however the result is acceptable and it is fair to say a linear relationship exists.
5.4.2. Checking the necessary assumptions

5.4.2.1 Individual linear dependency

The assumption that “the relationship between the response variable $y$ and the regressors $x$ is at least approximately linear” can be investigated by constructing a scatter plot of the quantitative regressors of the final model *area, monthly fee, and floor* against the response variable *price*.

![Plot of living area against the sold prices](image)

Plot 5.9: The living area plotted against the selling price

As can be seen in the plot above, where a regression with the single regressor *living area* is done against the response variable *selling price*, the data set follows a clear trend. The living area can be said to carry and approximately linear relationship with the selling price, which also is a very intuitive statement.
Plot 5.10: The monthly fee plotted against the selling price

Plotting the monthly fee against the selling price shows a near linear dependency between the regressor and the response variable, although it is not as clear as the previous one where the living area showed a clear linear trend. This difference can be the result of many difference factors. One of which could be that the monthly fee does not correlate to the selling price as much as the living area does, since some condos have very low, or almost zero, fees per month since the housing association sometimes does not require money from the owner, or requires a very low sum even though the condo in itself perhaps had a relatively high price. This in its turn can be due to the fact that associations sometimes are independent and sufficiently wealthy so the contribution from members is not needed as an economical factor.

Plot 5.11: The floor plotted against the selling price
Observing the plot above, it can be seen that the majority of the observations are distributed between 0 and 5 floors – which is the interval in which the observations follow a more clear linear trend. Meaning that if one examines the tendency of the data between \(0 < x < 5\), it seem to show higher values for \(y\) as the value for \(x\) progresses – a positive linear relationship. A more sporadic and random relationship between the selling price and floor seems to show up, when observing floor numbers higher than 5. It is important to point out that using dummy variables for the floor number would have been a more eligible method for evaluating this correlation. Intuitively, a penthouse floor – no matter the floor number – is expected to cost considerably more than any other floor in the same building, and the ground floor less. Unfortunately this could not be performed since the data obtained from Booli’s API did not deliver information regarding how many floors the building in which the condo is situated in had in total. Hence, the floor could not be set in relation to the total number of floors and the data did not carry any more significance than what the floor number was. Elaborating further, since the buildings in Stockholm do not all have the same number of floors, the data understandably does not show a clear linear trend.

Although as mentioned before, it is easier to spot a linear relationship between \(x\) and \(y\) when looking at the interval \(0 < x < 5\), this is mathematically backed up since the majority of the data used in this project had floors between 0 and 5. The number of observations with a floor number between 0 and 5 was 6640, given the data set of 8040 observations – this is a clear majority. Furthermore, the data points that can be seen on the right-half of the graph – the observations with high floor numbers but considerably low selling price – were noted to belong to the districts that the final model predicted to be cheaper, e.g. postal code 118. These districts, that also tend to lie closer to the borders of Stockholm city, have a higher number of residential areas, and thus have a considerably higher number of tall apartment buildings.

### 5.4.2.2 Error mean, variance, normality and correlation

The assumptions from section 2.1.1 state that:

1. The expected error term should have a mean of zero.
2. The error terms should have the same variance.
3. The errors should be uncorrelated.
4. The errors should be normally distributed.
The first and last assumption presented in this section are evaluated using the sample data, previously used for the out of sample validation. For the error term to have a mean of zero, it is implied that the predicted prices are bordering on the actual prices. Since the acquired model is not perfect, naturally, the mean error will also not be perfect – i.e. zero. The obtained mean error of the sample set was 18 677, which is considerably low in relation to the high condominium prices. Instead, the normal distribution of the errors can be checked in hopes of the residuals being normally distributed around zero.

The normal distribution of errors are presented below though a histogram and normal Q-Q plot constructed for the errors, with the errors being defined as the difference between the actual prices and the predicted prices. The clear normality of the sample errors is shown through the Q-Q plot, and the histogram shows the normality density around zero.

![Normal Q-Q plot of errors and Histogram of residual density](image)

Plot 5.12: A normal Q-Q plot of the errors of the sample set, and a histogram of residual density against the normal density

To examine assumption two and three – “the error terms should have the same variance” and “the error terms should be uncorrelated” – heteroscedasticity can be checked, since the detection of heteroscedasticity would be a violation of the assumptions. Heteroscedasticity can be detected both graphically and statistically. Graphically, possible heteroscedasticity of the sample fit can be detected through the plot presented below.
Plot 5.13: A plot of residuals against the fitted values, and a plot of scale location

If there is absolutely no heteroscedasticity present in the model, the data points in the plots should be scattered randomly and equally around the x-axis. The red line shown in the plots should thus be a straight line [43]. This means that there is some heteroscedasticity detected in the acquired final model, as there is a downwards curve towards the higher fitter values.

As earlier mentioned, since the acquired model is not perfect, it is unreasonable to also expect perfect homoscedasticity. A remedy to lessening the heteroscedasticity in this case would be the removal of the influential points or outliers. A simple test with the function `outlierTest` in R shows that the most extreme outliers of the data – i.e. the data with a higher or lower studentized residual value than ideal – are the same deviating points as in the plot above.

Statistically, the heteroscedasticity can be detected by conducting a non-constant variance test. The table below shows the p-value obtained through a Breusch-Pagan test of the fit, also conducted in R.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 21.23211, Df = 1, p = 4.0689e-06
```

Table 5.8: The values obtained through a Breusch-Pagan test

As mentioned in section 2.1.2.3, the heteroscedasticity in the fit can be considered insignificant since the obtained p-value is less than 0.05.
6. Conclusion

With the confidence interval being 95% in this study, the final model presented in section 4.1 with the specified regressor variables would be able to forecast the selling price of most apartments in Stockholm city with a ± 2.5% error margin. Note that the model can only be applied on the districts of Stockholm with postal codes between 112-118, since these ones are included as regressors in the model.

Observing the generated model, the most important variable is the living area, the next most influential variable is the floor, the third is the monthly fee, the remaining variables are dummies and have separately extensive impact on the response variable, the selling price. The data set on which the regression fit was conducted was achieved from the database from Booli and filtered down to a total of 8040 data points. To further improve the predictive ability of this model, an even bigger set of data, additional regressors and even more thorough analysis of outliers and variables is desirable.
7. Bibliography

5. Mail conversation with Mikaela Lagercrantz, developer at Booli. retrieved 2019-04-04
6. Yale University, Department of Statistics and Data Science (date not specified), *Multiple Linear Regression*, accessed 2019-20-03, <http://www.stat.yale.edu>
   <https://sv.wikipedia.org>
   <https://v8doc.sas.com>
19. ibid
   <https://v8doc.sas.com>
   <https://stattrek.com>
   <https://sv.wikipedia.org>
27. Minitab Blog Editor (2013), Enough Is Enough! Handling Multicollinearity in Regression Analysis, accessed 2019-05-02,
   <https://blog.minitab.com>
30. ibid, p. 351
   <https://www.statisticssolutions.com>
   <https://sv.wikipedia.org>


40. Gustafsson, A., Wogenius, S. (2014), *Modelling Apartment Prices with the Multiple Linear Regression Model*, Degree project in applied mathematics and industrial economics, Royal Institute of Technology

41. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013) *An Introduction to Statistical Learning with Applications in R*, Springer, New York


43. Snee, R. (1977), *Validation of Regression Models: Methods and Examples*, p. 418
