Stockholms
universitet

Prediction of Apartment Prices at Södermalm - A Regression Analysis

Olivia Lundberg

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:30
Matematisk statistik
December 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2015:30**
http://www.math.su.se

Prediction of Apartment Prices at Södermalm - A Regression Analysis

Olivia Lundberg[*]

December 2015

Abstract

The main aim of this study is to investigate the factors that have a potential influence on the final selling price of apartments. The use of multiple linear regression and a number of transformations will result in finding an informative model that describes if and how the predictor variables studied influence the response variable of final selling price of apartments. The predictive ability of the models chosen will be investigated in the hope of using the final model for future predictive purposes of selling prices. The study has been limited to investigating the area of Södermalm in Stockholm, Sweden and the final conclusions can be applied to this area only. The final model that best fit the data included a use of log transformations on the response variable and one of the predictor variables. The variable with the most influence on the selling price of apartments was, as suspected, the area of the apartment. Variables that showed non-significant within the final model were regarding the brokerage company used as well as during which season the apartment was listed on the market.

[*] Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: olivia.lundberg@gmail.com. Supervisor: Mathias Lindholm.

Abstract

The main aim of this study is to investigate the factors that have a potential influence on the final selling price of apartments. The use of multiple linear regression and a number of transformations will result in finding an informative model that describes if and how the predictor variables studied influence the response variable of final selling price of apartments. The predictive ability of the models chosen will be investigated in the hope of using the final model for future predictive purposes of selling prices. The study has been limited to investigating the area of Södermalm in Stockholm, Sweden and the final conclusions can be applied to this area only. The final model that best fit the data included a use of log transformations on the response variable and one of the predictor variables. The variable with the most influence on the selling price of apartments was, as suspected, the area of the apartment. Variables that showed non-significant within the final model were regarding the brokerage company used as well as during which season the apartment was listed on the market.

Preface and Acknowledgements

The following thesis represents the final 15hp in the Bachelor's program in Mathematics and Economics within the department of Mathematical Statistics at Stockholm University and will result in a Bachelor of Arts.

I would like to thank my supervisor Mathias Lindholm at Stockholm University, for mentoring me through the thesis. I would also like to give thanks to my former classmate Björn Smedberg for helping me with obtaining data, as well as my parents and fellow classmates for supporting me through the entire program.

Table of Contents

Abbreviations

AIC	Akaike Information Criterion
VIF	Variance Inflation Factor
i.i.d	Independent and Identically Distributed
BLUE	Best Linear Unbiased Estimate
MSEP	Mean Square Error of Prediction
RMSEP	Root Mean Square Error of Prediction
TSS	Total Sum of Squares
ESS	Explained Sum of Squares
RSS	Residual Sum of Squares
df	Degrees of Freedom

1 Introduction

An interest in finding out the reason behind the high selling prices of certain apartments in Stockholm is the main reason behind this thesis. A wish to gain more knowledge of the Swedish real estate market, specifically at Södermalm, as this is a part of Stockholm that seems to have grown in popularity in the past few years. The real estate market in inner-city Stockholm is well known amongst the Swedish population for being expensive, and the prices have been increasing for the past few years. "Apartment prices on Södermalm has risen by an average of 21 percent in the past year. During the same period, consumer price inflation rose by 0.1 percent."[1]

To do this, an in-depth regression analysis will be pursued to find a model which correctly describes the relationship between the selling price of an apartment and a number of explanatory variables. A use of linear models, log and Box-Cox transformations and model selection devices such as stepwise selection will result in concluding which variables have a significant effect on the selling price of apartments. The final chosen model will then be used for prediction of future apartment prices, in order to assess the prediction ability.

This study will be helpful for people who are interested in purchasing or selling an apartment in the south part of inner-city Stockholm. It will be easier for both private parties as well as brokers to know what to look for when an apartment enters the market. Questions for the reader to keep in mind throughout the thesis, that can be considered as the focal point of this study, are stated below.

- Is it, as expected, the area of the apartment that has the most influence on the selling price, or are other factors of more importance?
- Can some factors be disregarded in the pricing of an apartment?
- Will the final model be able to be used for predicting the future selling price of apartments?

The thesis is divided into five chapters, starting out with the *Introduction* where the background of the subject has been discussed and the purpose of the analysis was stated. The *Theoretical Framework* will provide knowledge of the theory supporting the methods used in the following *Analysis and Results* chapter. Finally, all results will be discussed and conclusions drawn in the *Discussion* chapter.

2 Theoretical Framework

Within the following section, the necessary theoretical material that will be referred to and utilized in the remainder of the thesis will be reviewed. The reader will receive information crucial to understanding the analysis following as well as for discussing the results. *Lineära Statistiska Modeller* by Rolf Sundberg [2] as well as *Practical Regression and Anova using R* by Julian Faraway [3] will be used as references throughout this chapter.

2.1 Linear Regression

There are a number of assumptions when using linear regression, stated below:

Linearity : the relationship between the response variable and the predictor variables are assumed to be linear to each other

Normality : the residuals of the model are assumed to be normally distributed

Homoscedasticity : constant variance of residuals

Absence of Multicollinearity : predictor variables are not dependent on each other

2.1.1 Simple Regression

The model for simple linear regression is defined as the following:

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, ..., n \tag{2.1}$$

where $\epsilon \sim N(0, \sigma^2)$ and i.i.d. in (2.1)

2.1.2 Multiple Regression

The model for multiple linear regression is defined below:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_m x_{mi} + \epsilon_i, \qquad i = 1, ..., n \tag{2.2}$$

where $\epsilon \sim N(0, \sigma^2)$ and i.i.d.

Written into matrix form, we have the following:

$$y = X\beta + \epsilon, \tag{2.3}$$

where $y = (y_1 ... y_n)^T, \epsilon = (\epsilon_1 ... \epsilon_n)^T, \beta = (\beta_1 ... \beta_n)^T$ and

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \multicolumn{5}{c}{\dotfill} \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

where the ones account for the intercept parameter α.

2.1.3 Hypothesis Testing

The hypothesis we will be testing is whether the parameter term β has a significant effect on the response variable.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_n = 0$$

2.1.4 Estimating Parameters

The method of parameter estimation used is ordinary least squares (OLS) using a built-in system in R. If the errors are correctly assumed i.i.d, OLS returns $\hat{\beta}$ as the maximum likelihood estimator, and is according to the Gauss-Markov theorem the best linear unbiased estimator (BLUE).

The best estimate of β is the one which minimizes the sum of the squared errors, $\epsilon^T \epsilon$. With some calculations, we get that $\hat{\beta} = (X^T X)^{-1} X^T y$, where $\hat{\beta}$ is normally distributed with $N(\beta, \sigma^2 (X^T X)^{-1})$.

This is plugged into $\hat{y} = X\hat{\beta}$ which in turn provides the estimation for the residuals. This is used for calculation of the estimate for σ^2:

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n - m - 1} = \frac{\sum_1^n (y - \hat{y})^2}{n - m - 1} \tag{2.4}$$

where $n - m - 1$ are the degrees of freedom (df).

Finding confidence intervals for the parameters is done by the following:

$$\hat{\beta}_i \pm t_{(\alpha/2)}(n - m - 1)\hat{\sigma}\sqrt{(X^T X)^{-1}} \tag{2.5}$$

where (n-m-1) are the df. The t-test is used in this which is also used for testing the hypothesis that β has a significant effect on the response.

2.2 Model Efficiency

Several things need to be considered when determining the efficiency of the model you have chosen. The ways used in this analysis are explained below.

2.2.1 Akaike Information Criterion

The Akaike Information Criterion (AIC) measures the quality of models for a given data set. The preferred model will be the one with the lowest AIC value. It is calculated using the log likelihood within the formula:

$$AIC = 2k - 2logL(\hat{\beta}|y) \tag{2.6}$$

2.2.2 Multicollinearity

Multicollinearity can be the results of, for instance, having two similar parameters that in turn are highly correlated to each other. ”Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.” [4]

A way of finding issues with multicollinearity within data is investigating the VIF values within the model.

2.2.3 Variance Inflation Factor

The Variance Inflation Factor (VIF) calculates how much the variance of the estimated regression coefficients $\hat{\beta}_i$ is increased when combined with parameters with high collinearity. It is a way of deciding whether multicollinearity will show to be an issue in the future analysis. The VIF value that is used as a maximum limit is usually 5 or 10, this is when multicollinearity will be considered a problem. VIF is defined as

$$VIF = \frac{1}{1 - R_i^2} \tag{2.7}$$

where R_i^2 is the coefficient of determination for X_i with the remaining predictor variables on the right hand side.

2.2.4 Coefficient of Determination

The coefficient of determination, R^2, is the percentage of variance explained by the model. The definition of this is:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \overline{y})^2} \tag{2.8}$$

where ESS=Explained Sum of Squares, RSS=Residual Sum of Squares, TSS= Total Sum of Squares, and TSS=ESS+RSS.

A R^2 value of 1 indicates that 100% of the variation in the response variable can be explained by the variation in the model. The Adjusted R^2 accounts for the amount of predictor variables included in the model and calculates the decrease in residuals, having the following relationship with R^2:

$$1 - R_{adj}^2 = (1 - R^2)\frac{df_{tot}}{df_{res}} \tag{2.9}$$

2.2.5 Stepwise Regression

Within stepwise regression there are three main ways of going about choosing a final model. In R, the inclusion of the parameter in the model relies on the AIC value.

- Forward Selection : An empty model will analyse every variable and include those that are significant.

- Backward Elimination : A full model will be checked step by step to see if there are variables that need to be excluded from the model.
- Stepwise Selection : After each instance of adding a variable by forward selection, all previous variables added are checked using backward elimination.

We will be using stepwise selection as the main stepwise regression tool, as it uses a combination of the forward selection and backward elimination.

2.2.6 Residuals

The residuals can be calculated from equation (2.2).

$$\hat{\epsilon} = y - \hat{y}$$

One important diagnostics plot is the one which plots the residuals $\hat{\epsilon}$ against the fitted values $\hat{y}$. This is where non-constant variance can be spotted. For homoscedastic data, there will be a band of residuals symmetrically distributed around the x-axis. As the assumption of regression is that the residuals are normally distributed, it is also important to check whether there is any skewness of the residuals.

”Skewness is a measure of the degree of asymmetry of a distribution. If the left tail (tail at small end of the distribution) is more pronounced than the right tail (tail at the large end of the distribution), the function is said to have negative skewness. If the reverse is true, it has positive skewness. If the two are equal, it has zero skewness.” [5]

The formula for calculating the Fisher-Pearson coefficient of skewness is defined as the following:

$$g_1 = \sum_{i}^{n} \frac{(y_i - \overline{y})^3/n}{s^3} \tag{2.10}$$

where s is the standard deviation.

2.3 Model Transformation

In order to improve fit of models and possibly correct disturbances such as heteroscedasticity, transformations of the variables may be of help. Faraway discusses this in chapter 8 in 'Practical Regression and Anova using R'.

2.3.1 Log Transformation

In the typical multiple linear model shown in Equation (2.2), the errors enter additively to the model.

$$logy_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_m x_{mi} + \epsilon_i, \tag{2.11}$$

When taking the log of the response, the model with the original scale of the response changes. The errors enter multiplicatively in to the model, which means that interpreting the regression coefficients is different when having transformed the response variable. In this case, a one unit increase in x_1 would result in a multiplicative increase of $e^{\hat{\beta}_1}$ on the response.

$$\hat{y}_i = e^{\hat{\alpha}} e^{\hat{\beta}_1 x_{1i}} e^{\hat{\beta}_2 x_{2i}} .... e^{\hat{\beta}_m x_{mi}} \tag{2.12}$$

Other possible transformations would be to instead transform one or more predictor variables (2.13) or transforming both response and predictor variables (2.14).

$$y_i = \alpha + \beta_1 logx_{1i} + \beta_2 logx_{2i} + ... + \beta_m logx_{mi} + \epsilon_i, \tag{2.13}$$

$$logy_i = \alpha + \beta_1 logx_{1i} + \beta_2 logx_{2i} + ... + \beta_m logx_{mi} + \epsilon_i, \tag{2.14}$$

2.3.2 Box-Cox Transformation

The Box-Cox method, explained in chapter 9 in 'Practical Regression and Anova using R'[3] is designed for choosing the transformation to best fit the data available. Transforming the response variable is done in the following way:

$$t_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \text{logy} & \lambda = 0 \end{cases} \tag{2.15}$$

The value for $\hat{\lambda}$ is found by maximizing the profile log-likelihood shown in the equation below.

$$L(\lambda) = \frac{n}{2} log(\frac{RSS_\lambda}{n}) + (\lambda - 1) \sum_i^n log(y_i) \tag{2.16}$$

where RSS_λ is the residual sum of squares when $t_\lambda(y)$ is the response variable.

2.4 Prediction

Testing the fit of a model can be done by using it on a new set of historical data. This is a sort of backtesting done by using the already estimated beta values on the new data, to achieve a vector of new response values.

$$y_i^* = X_{new} * \hat{\beta} \tag{2.17}$$

In order to compare one would look at the residuals $y_i - y_i^*$ where y_i is the actual historical values one can compare the predicted values with. It is necessary to be careful when calculating the new predicted values for the transformed models. For example, predicting for a log model mean having to transform back $e^{y_i^*}$.

2.4.1 Mean Square Error of Prediction

If $y_i{}^*$ is a vector of n predictions, and y_i is the vector of observed values corresponding to the inputs to the function which generated the predictions, then the mean square error of the prediction can be estimated by

$$\text{MSEP} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2 \tag{2.18}$$

The square root of this measurement can also be used as a prediction of the size of the errors, called RMSEP (Root Mean Square Error of Prediction).

3 Data

The data used in this thesis has been obtained through the company Booli Search Technologies AB [6] which is an information search service based on the Swedish real estate market. Data was downloaded via the web site's own API and include observations from the real estate market in Stockholm since year 2012, with a total of 6281 apartments sold in the region of Södermalm. The language used for the statistical computing in this analysis was R, with Excel used as a stepping stone for data import.

3.1 Original Data Set

The original data file obtained included the following variables:

PriceS : The final selling price of the apartment.

PriceL : Accepted price when listing first came on the market.

Area : The size of the apartment measured in square meters.

Rent : Monthly cost of living in said apartment.

Year : Construction year of apartment building.

DateL : The date when the apartment was listed.

DateS : The date when the apartment was sold.

Region : The specific region within Södermalm.

Floor : Which floor the apartment is on.

Broker : The brokerage company used for the sale.

Ocean : Distance to the ocean.

3.2 Variable Transformation

Removing, adding or changing variables in regards to our original data set can help simplify and prevent issues when doing the analysis. Something to keep in mind is that we plan on using the final model for prediction, which prevents us from using predictor variables that contain information on already sold apartments. If, for example, one would like to find out whether the amount of days that the apartment was out on the market before it was sold, it could easily be added to the model. However, since the model will be used for prediction, one will not have access to the selling date ’DateS’, which is why it will be excluded from the analysis. For prediction purposes, data was separated at the 2014/2015 year mark within the ’DateL’ variable. The reason for this is so that we can fit a model on our data up until 2015, and then use this model to assess the fit on our leftover data in year 2015.

At first glance, it could be seen that there were around 1500 missing observations within the ’Floor’ variable, some of which were included in rows that had several missing parameter values. In order to see if there was any systematic reason for these, we took a look at the individual observations for these rows and plotted them against ’PriceS’, which can be seen in Figure 1 in the appendix. We concluded that there was no specific reason for this and in the decision of either excluding the entire ’Floor’ variable from the analysis, or removing all rows that contained missing observations, we decided on the latter.

The data obtained from Booli regarding in which regions of Södermalm the apartments were located were not correctly specified. Using longitude data for each listing, we redefined the variable ’Region’ using three overall regions. Two other variables that were grouped into categorical variables were ’Broker’ and ’Floor’. ’Season’ was created as a new categorical variable that used the information within ’DateL’ to specify during which time of the year the apartment was released on the market. The continuous variable ’Year’ was changed to $'Age' = 2015 -' Year'$, a variable showing how old the apartment building is.

Considering the fact that highly correlated predictor variables create issues in regression models, we decided to create two new variables that used the information from variables that had a higher probability of creating problems. 'PriceL' is quite possibly very highly correlated with our response variable, which is the reason for the creation of $'PriceArea' = \frac{'Price'}{'Area'}$. The variables 'Rooms' and 'Area' are also quite possibly highly correlated, resulting in the creation of a variable showing the area per room, $'AreaRoom' = \frac{'Area'}{'Rooms'}$.

3.3 Final Data Set

After the modifications in our given data set, we present the final collection of variables that will be used in our analysis.

PriceS : Continuous response variable, ranging from 1450-15200($1000kr$).

PriceArea : Continuous variable, ranging from 34782-115122(kr/m^2).

AreaRoom : Continuous variable, ranging from 13.5-54($m^2/room$).

Area : Continuous variable, ranging from 11-217(m^2).

Rent : Continuous variable, ranging from 100-8664(kr).

Age : Continuous variable, ranging from 4-370($years$).

DateL : Continuous variable, ranging from 2011-10-21 to 2014-12-31($date$)

Region : Categorical variable, observations 'Center'($baseline$), 'East' and 'West'.

Floor : Categorical variable, observations 'Low'($baseline$), 'One', 'Two', 'Three', 'Four', 'High'.

Broker : Categorical variable, observations 'Small'($baseline$), 'Medium', 'Large'.

Season : Categorical variable, observations 'Autumn'($baseline$), 'Winter', 'Spring', 'Summer'.

The response variable PriceS is plotted against the continuous variables in Figure 1 and the categorical variables in Figure 2, in order to get an overview.

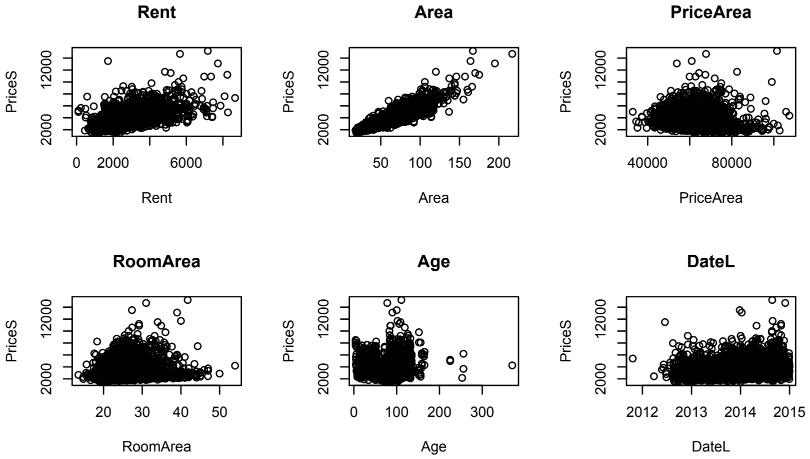

Figure 1: Plot of Response vs Continuous Variables

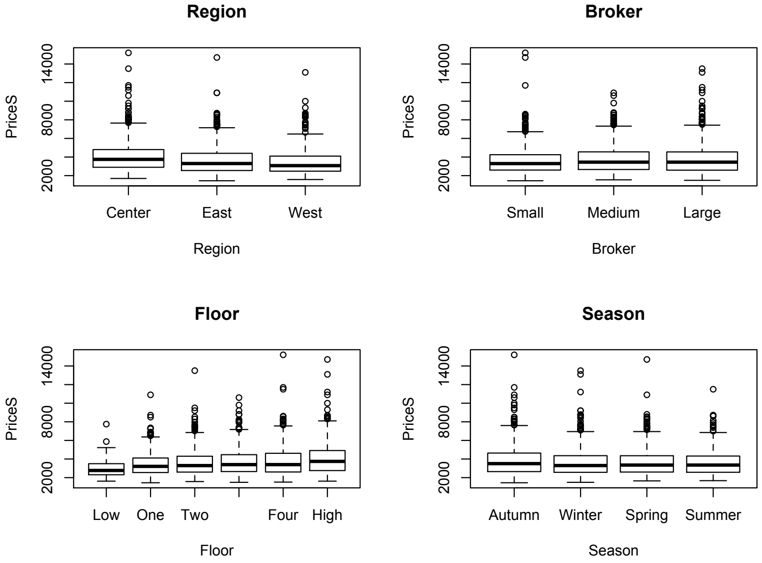

Figure 2: Plot of Response vs Categorical Variables

4 Analysis and Results

This section shows the analysis technique used to find a model that fits our data. Analysed below will be regarding each section of the theoretical framework.

4.1 Linear Regression

In order to get an overview of how well the predictors work on their own against the response variable, we perform simple linear regression of each predictor variable, with the results shown below:

Variable	P-value	R^2
Region	$< 2.2e - 16$	0.02874766
Rent	$< 2.2e - 16$	0.4164523
Floor	$< 2.2e - 16$	0.02418968
PriceArea	$9.45e - 14$	0.01622919
AreaRoom	0.04185	0.001574889
Area	$< 2.2e - 16$	0.8236692
Broker	0.2196	0.002789041
Age	0.1011	0.0003346439
Season	0.0001737	0.002667207
DateL	$< 2.2e - 16$	0.01484966

Table 1: Linear Regression Values

It is shown above that all variables except 'Broker' and 'Age' were significant on the 0.05 level in the simple regression. The variable that seems to have the most influence on the response is 'Area' with a R^2 value of 0.82.

4.2 Multiple Linear Regression

As discussed shortly in the *Data* chapter, the reason for exchanging the original predictor variable 'Rooms' for the new 'AreaRoom' will be discussed first. When attempting to place both variables in the full multiple linear regression model, we checked the correlation between them, which came out to 0.912. Running the regression made it evident that the VIF values of both of these variables were above 5. We figured this might be a reason for the heteroscedasticity evident in the model, which can be seen in the residual plot in Figure 3 below.

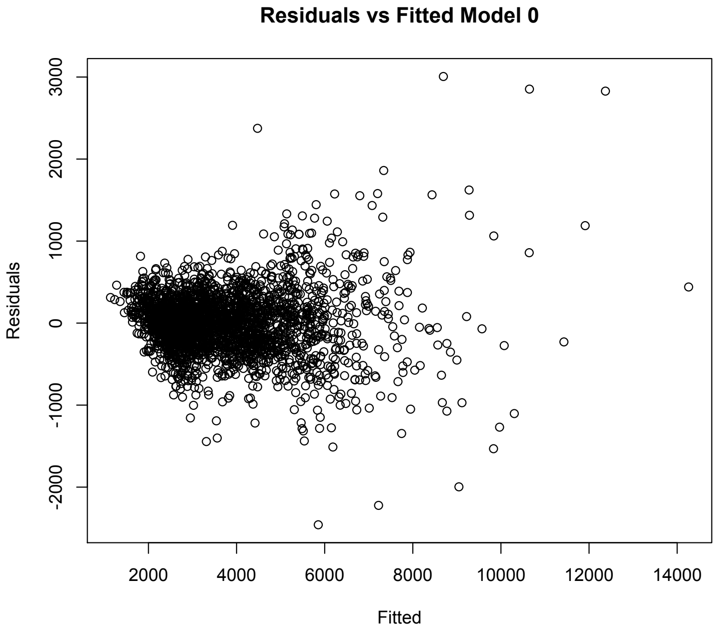

Figure 3: Residuals vs Fitted Model 0

In an attempt to solve this problem, we made the changes in variables and equation (4.1) below became our so-called ”original” model.

$$PriceS_i = \alpha + \beta_1 PriceArea + \beta_2 AreaRoom + \beta_3 Area + \beta_4 Broker + \\ + \beta_5 Age + \beta_6 DateL + \beta_7 Region + \beta_8 Rent + \beta_9 Season + \beta_{10} Floor + \epsilon_i \tag{4.1}$$

In this model, the VIF values had lowered and were all below the critical value of 5. A necessity when determining the fit of a model is checking data for outliers. In Figure 4 below we can see that there are a few observations that are considered outliers.

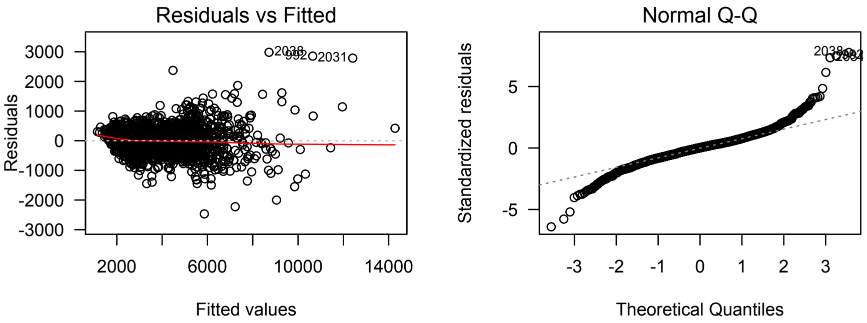

Figure 4: Full Model Residual Plots

We wish to see what changes occur when removing the three outliers with the highest absolute residual values. These three observations also had the highest Cook’s D values in the data set, which is an estimate for the influence of the observation, giving us reason for removing them as outliers. This lowered skewness of the residuals greatly, as it went from 0.3972 to 0.0721.

However, when looking at the residuals in Figure 5 below, there was still heteroscedasticity evident and the assumption of normality of the residuals does not seem to be correct.

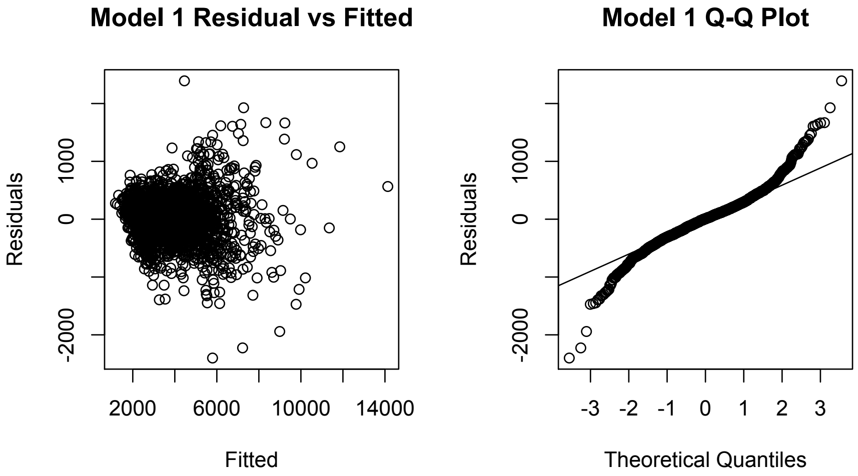

Figure 5: Model 1 Residual Plots

In order to try to even out out residuals, we will try a few of transformations. First out is simply taking the log transformation of the response variable PriceS, now becoming logPriceS, but keeping the rest of the variables the same.

To compare these models, we can analyse Figure 6 below. There is a definite change in the residuals, but heteroscedasticity does still seem to be evident. However, we can see in the Q-Q plot that the residuals are starting to look more normally distributed.

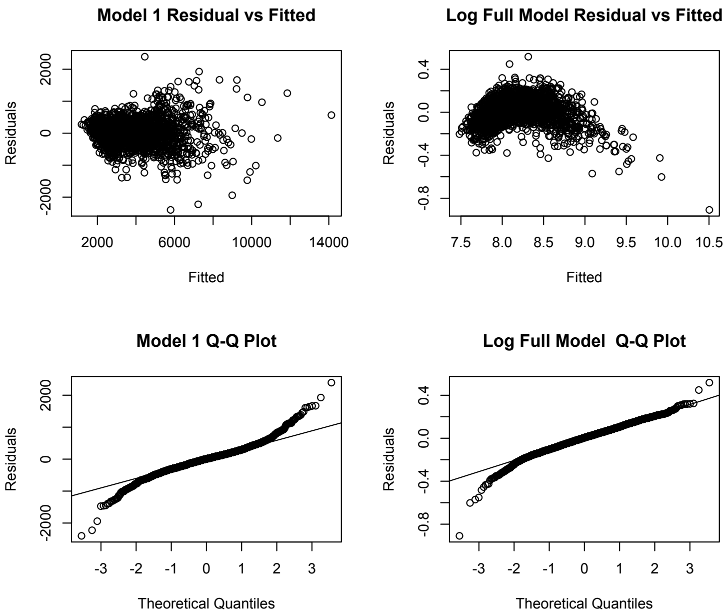

Figure 6: Model 1 vs Log Full Model Residuals

Since it looks like we are on the right track, we will do some further transformations of some of the explanatory variables. Keeping the response as logPriceS, we will now focus on the explanatory variables. One possibility would be to log transform our variable 'Area', which is probably one of the variables with the most influence on our response (considering the output in Table 1).

We will be using a version of Equation (2.14) where we transform the response and only one of the predictor variables. Our new Log Model is defined in equation (4.2) below.

$$\begin{aligned} logPriceS_i = \alpha + \beta_1 PriceArea + \beta_2 AreaRoom + \beta_3 logArea + \beta_4 Broker + \\ + \beta_5 Age + \beta_6 DateL + \beta_7 Region + \beta_8 Rent + \beta_9 Season + \beta_{10} Floor + \epsilon_i \end{aligned} \tag{4.2}$$

We now look at how the residuals of the model have changed after the transformation.

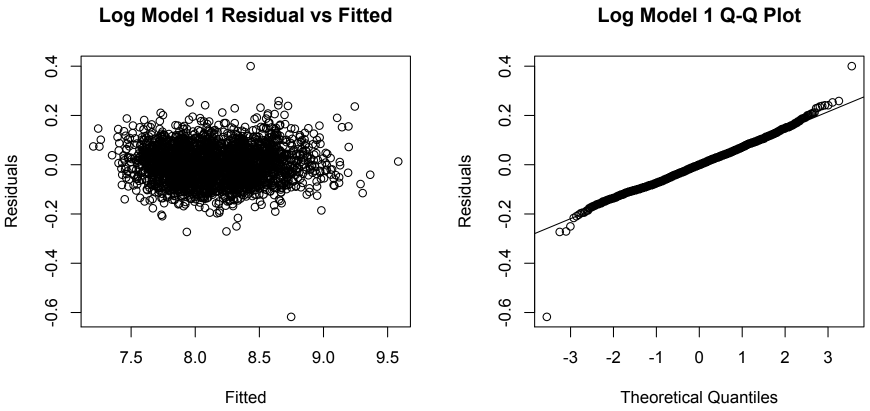

Figure 7: Log Model 1 Residuals

In Figure 7 we can see that there are great changes in the residuals. They have levelled out and seem more homoscedastic and according to the Q-Q plot, the assumption that the residuals are normally distributed seem correct.

We decided to try out an additional transformation that could prove profitable, a Box-Cox transformation of our response variable. The plot below shows us where the λ value in Equation (2.16) is maximized, which in our case was 0.6667.

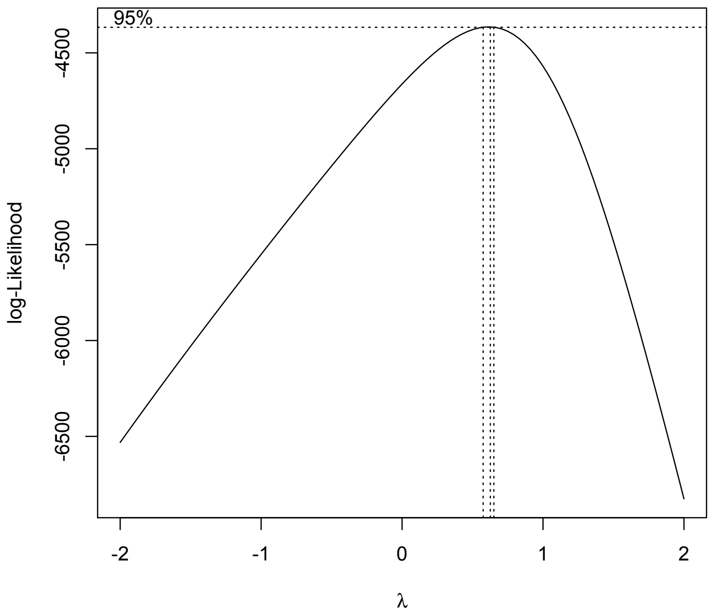

Figure 8: Box-Cox Plot

Plugging this in to the Box-Cox transformation formula, we get our new transformed response variable, boxPriceS:

$$t_\lambda(y_i) = \frac{y_i^{0.6667} - 1}{0.6667} \tag{4.3}$$

Our Box-Cox Model is therefore defined as:

$$
\begin{aligned}
boxPriceS_i = \alpha + \beta_1 PriceArea + \beta_2 AreaRoom + \beta_3 Area + \beta_4 Broker + \\
+ \beta_5 Age + \beta_6 DateL + \beta_7 Region + \beta_8 Rent + \beta_9 Season + \beta_{10} Floor + \epsilon_i
\end{aligned}
\tag{4.4}
$$

Looking at the residuals in Figure 9, the transformation of the response does not seem to solve the problems regarding heteroscedasticity.

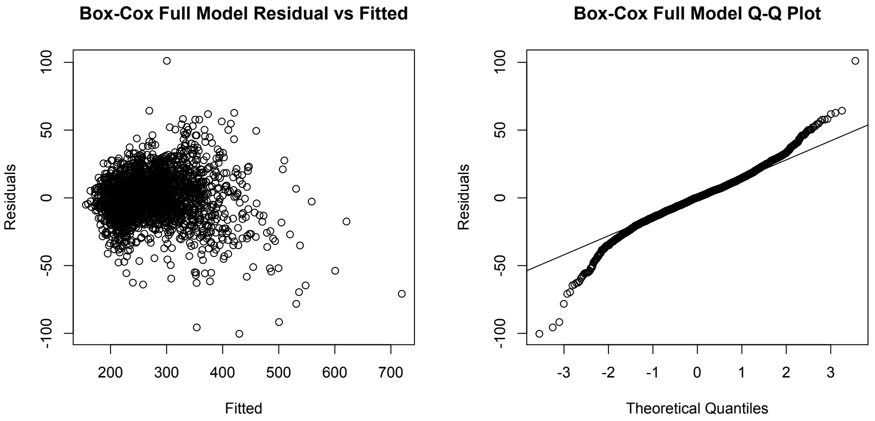

Figure 9: Box-Cox Full Model Residuals

In order to check the models for variables that might not be significant, we will be using stepwise selection.

4.2.1 Stepwise Selection

In this section we will be using stepwise selection for the different models in the previous section. In equations (4.5)-(4.7) below you can see the final four reduced models which now only include significant variables. The output from R showing the coefficients of all models can be seen in Table 4-6 in the appendix.

Model 1

$$\begin{aligned} PriceS_i &= \alpha + \beta_1 PriceArea + \beta_3 Area + \beta_5 Age + \\ &+ \beta_6 DateL + \beta_8 Rent + \beta_9 Season + \beta_{10} Floor + \epsilon_i \end{aligned} \tag{4.5}$$

Log Model 1

$$\begin{aligned} logPriceS_i &= \alpha + \beta_1 PriceArea + \beta_2 AreaRoom + \beta_3 logArea + \\ &+ \beta_5 Age + \beta_6 DateL + \beta_8 Rent + \beta_9 Season + \beta_{10} Floor + \epsilon_i \end{aligned} \tag{4.6}$$

Box-Cox Model 1

$$\begin{aligned} &boxPriceS_i = \alpha + \beta_1 PriceArea + \beta_2 AreaRoom + \beta_3 Area + \\ &+ \beta_6 Age + \beta_7 DateL + \beta_8 Region + \beta_9 Rent + \beta_{10} Season + \beta_{11} Floor + \epsilon_i \end{aligned} \tag{4.7}$$

A comparison between important values in the full models and the reduced ones can be made in the following table.

	Adj R^2	AIC	Skewness	Residual SE
Model 1	0.934335	31142.2	0.07211903	373.8
Step Model 1	0.9343389	31138.07	0.06670592	373.8
Log 1	0.9583472	-13700.21	-0.04677067	0.07345
Step Log 1	0.9583965	-13707.29	-0.04873061	0.0734
Box 1	0.9314929	16595.72	-0.1980825	23.45
Step Box 1	0.93155	16594.94	-0.2012799	23.46

Table 2: Regression Model Values

The decision of exchanging our original models with the reduced models and continuing the analysis using these was made. In order to further analyse the fit of the models, the residuals will be investigated and compared in between them. Below are Normal Q-Q plots of the residuals.

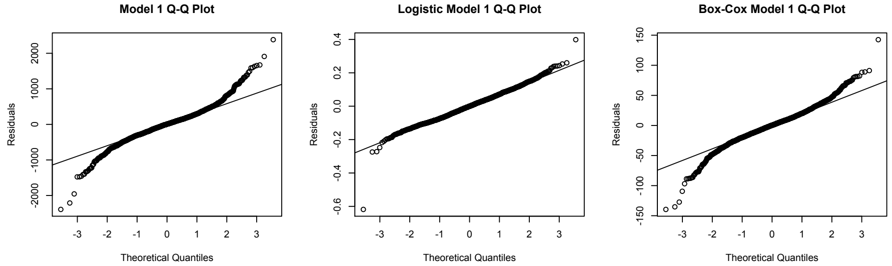

Figure 10: Q-Q Plot of Model Residuals

To further check the residuals of the models, we take a look at plots of the residuals versus the fitted values.

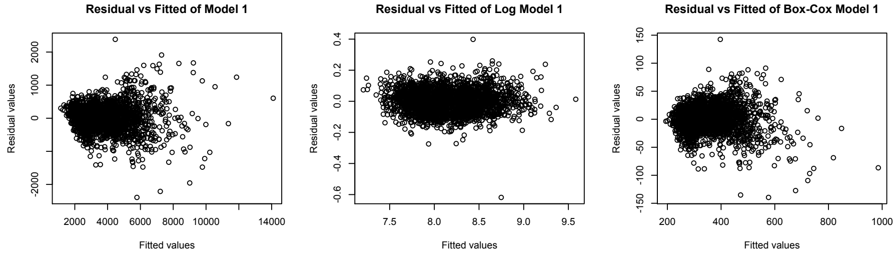

Figure 11: Residual vs Fitted of Models

Looking at the residuals of the three models and considering the overall results in Table 2, we can conclude that both transformed models are superior to the original model. It is also evident that Log Model 1 is the model that best fit our data, as the residuals seem to be normally distributed as well as being homoscedastic. It also has the highest $AdjR^2$ value, the lowest value for AIC, a residual skewness closest to zero as well as the lowest residual standard error.

4.3 Predictive Ability

Models (4.5)-(4.6) have been fitted using the data set up until 2014-12-31. In order to determine the predictive ability of these, the ”new” set of observations in year 2015 will be re-introduced to the analysis. We will be using the models fitted in the previous section on the new ”future” data set. Once these models have predicted the new values for PriceS, we will compare these to the actual values we have in the data set.

To get an overview of how well our models work, we can take a look at the actual values plotted against values that were predicted through our models with the new data set. If there would be a perfect fit, the predicted values would be equal to the actual values and the observations would be plotted as a straight line going through (0,1) in the following plot:

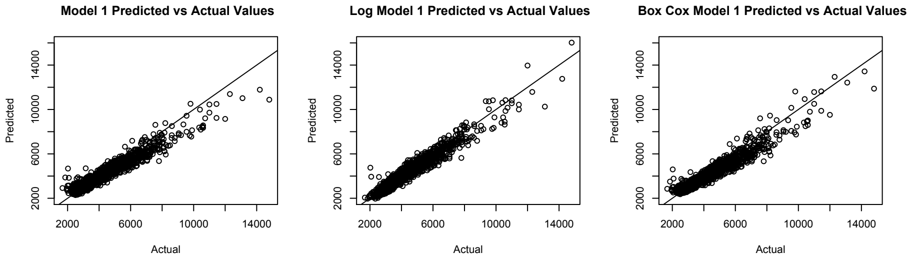

Figure 12: Actual vs Predicted Values

Next we look at the residuals of the actual values and the predicted values.

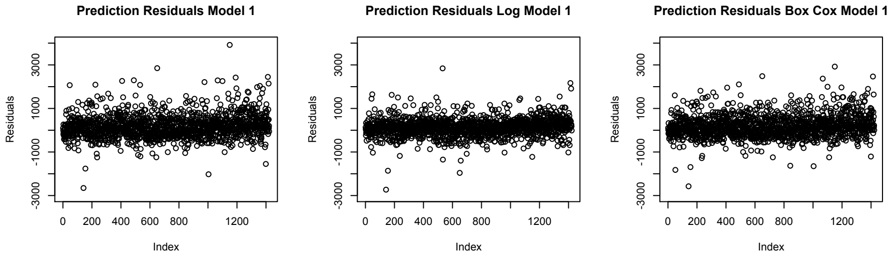

Figure 13: Prediction Price Difference

RMSEP values for each model are shown in the table below.

	MSEP Value	RMSEP Value
Model 1	311974.8	558.547
Log Model 1	179566.4	423.753
Box-Cox Model 1	294774.7	542.932

Table 3: Mean Square Error

We can see in the table above that the lowest value of MSEP and RMSEP in this case is for Log Model 1. What is necessary to remember is that all values have been calculated using the re-defined price of the actual price divided by 1000. Nevertheless, the Log Model 1 is proven superior in predictive ability.

5 Discussion

5.1 Results

It can be concluded that the preferred model is Log Model 1. The parameter values shown in Table 6 are from the regression for the response variable logPriceS. From the Log Model 1 equation (4.6) we therefore get the following multiplicative relationship for the response PriceS:

$$\begin{aligned} PriceS_i = e^{\alpha} * e^{\beta_1 PriceArea} * e^{\beta_2 AreaRoom} * e^{\beta_3 logArea} * \\ * e^{\beta_5 Age} * e^{\beta_6 DateL} * e^{\beta_8 Rent} * e^{\beta_9 Season} * e^{\beta_{10} Floor} * \hat{\epsilon}_i \end{aligned} \tag{5.1}$$

We will now have a closer look at the individual variables and their effect on 'PriceS'. The intercept value of $e^{1.17854187817} = 3.2496324$ would normally be interpreted as the selling price of the apartment when all predictor variables are set to zero. However, this is not realistic in this case. It is not possible that, for example, an apartment with an area of zero would enter the real estate market.

The β estimates for the significant variables included in the model were interpreted as follows:

PriceArea : A one unit change in PriceArea would result in a 0.00121% increase in PriceS as $e^{0.00001207461} = 1.0000121$.

AreaRoom : A one unit change in AreaRoom would result in a 0.1549% decrease in PriceS as $e^{-0.001549916996} = 0.9984520$.

logArea : The value 0.94084167204 means that a 1% increase in logArea creates a 0.94% increase in price.

Rent : A one unit change in Rent would result in a 0.00106% decrease in PriceS as $e^{-0.00001059427} = 0.9999894$.

Age : A one unit change in Age would result in a 0.05719% increase in PriceS as $e^{0.00057175970} = 1.0005719$.

DateL : A one unit change in DateL would result in a 0.01555% increase in PriceS as $e^{0.00015546334} = 1.0001555$.

Floor : As the baseline of this categorical variable is Low, all results will be in regards to that. If an apartment is on floor One instead of Low, 'PriceS' will increase by a multiple of $e^{0.00734908697} = 1.0073762$.
This interpretation is also used for the estimates for the floors
Two $e^{0.02594192111} = 1.0262813$, Three $e^{0.01913642331} = 1.0193207$, Four $e^{0.03361253183} = 1.0341838$ and High $e^{0.03741263613} = 1.0381213$.

Season : As the baseline of this categorical variable is Autumn, all results will be in regards to that. If an apartment is sold during Winter instead of Autumn, 'PriceS' will increase by a multiple of $e^{0.00609386304} = 1.0061125$.
This interpretation is also used for the estimates for Spring $e^{0.01075518973} = 1.0108132$ and Summer $e^{0.00562504232} = 1.0056409$.

Instead of stating the percentage changes in the response variable, one could interpret the estimates such that the selling price of the apartment would increase by a multiplicative factor of the estimate in question. One must keep in mind that all interpreted parameter values assume that all other predictor variables are held constant. Variables that were excluded from Log Model 1 in the stepwise selection because of insignificance were the following:

Region : This variable had an extremely high p-value of 0.96664 and was therefore not included in the final model.

Broker : This variable had a high p-value of 0.66067 and was therefore not included in the final model.

In conclusion, variables that have a significant positive effect on the final selling price of apartments 'PriceS' are 'PriceArea', 'logArea', 'Age', 'DateL', 'Floor' and 'Season'. Variables that have a significant negative effect on 'PriceS' are 'Rent' and 'AreaRoom'. Variables that does not have significant effects on the reponse are 'Region' and 'Broker'.

5.2 Limitations and Other Analyses

There were quite a few limitations in this analysis. The data used did not include several variables that most probably would have had an impact on the final selling price. A few examples of these types of variables are the existence of a balcony, an elevator, more specific regions of Södermalm as well as distance to the subway. These variables could also have an effect on the other predictor variables. An example of this would be 'Floor'. Living on the seventh floor without an elevator might not be the most popular apartment but with an elevator, it is definitely a better purchase and would result in a higher selling price. However, we were still able to find a model with a good fit with the data available.

"Final price information on condominiums is based on a structured automatic collection of final bids from open bidding on-line. This means that many end rates are included but not all. All brokerage firms do not present final prices openly on the site, and even though an agency usually does so, the seller can always choose not to display the bidding."[7] This quote shows that there are limitations to the data we obtained from Booli, as they do not obtain all listings nor possibly completely correct information because of the automatic collection of data. For future analyses, it is important to have data that is as correct and complete as possible, in order to prevent drawing any inaccurate conclusions.

There is some past existing knowledge regarding this area of analysis. There are other theses that have focused on this type of analysis and have results similar to the ones stated in this thesis. [8] For future analyses, one could use time series as well as checking more types of transformation of data, to see if this provides a model with a better fit. Another type of analysis that could be interesting to do is using the relative price as a response variable. This could show if for example brokers adjust the listed prices lower because they know for a fact from history that they will increase with a certain percentage. Is the concept of having an "accepted price" just a way of getting more people to the open houses, or is there a mutual understanding between brokers and buyers that the price will rise?

In checking predictive ability, using MSEP can possibly be misleading if one would use a data set in the future with contains observations that are very different to the ones used in the regression analysis.

References

[1] Anrell, O., and Bonnichsen, L. (2015). Här ökar priserna på borätter mest. Mitt I Södermalm, p. 4. Translated by author.

[2] Sundberg, R. (2014). Lineära Statistiska Modeller (2nd ed.). Stockholm: Department of Mathematics, Stockholm University.

[3] Faraway, J. (2002). Practical Regression and Anova using R. [Bath: University of Bath].

[4] Statistics Solutions,. (2015). Multicollinearity. Retrieved 7 November 2015, from https://www.statisticssolutions.com/multicollinearity/

[5] Weisstein, E. (2015). Skewness. Wolfram MathWorld. Retrieved 7 November 2015, from http://mathworld.wolfram.com/Skewness.html.

[6] Booli.se. Booli API.
Retrieved 29 September 2015, from http://www.booli.se/api.

[7] Support.booli.se. (2014). Visar Booli alla slutpriser?
Retrieved 23 September 2015, from http://support.booli.se/customer/en/portal/articles/126549-visar-booli-alla-slutpriser-.
Translated by author.

[8] Aguirre, C. (2014). Analys av lägenhetspriser i Hammarby Sjöstad med multipel linjär regression (Bachelor). Department of Mathematics, Stockholm Univeristy.

[9] Djurfeldt Djurfeldt, G., and Barmark, M. (2009). Statistisk Verktygslåda 2. Stockholm: Studentlitteratur.

[10] Laerd Statistics. (2015). How to perform a Multiple Regression Analysis in SPSS Statistics. Retrieved 5 October 2015, from statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php

6 Appendix

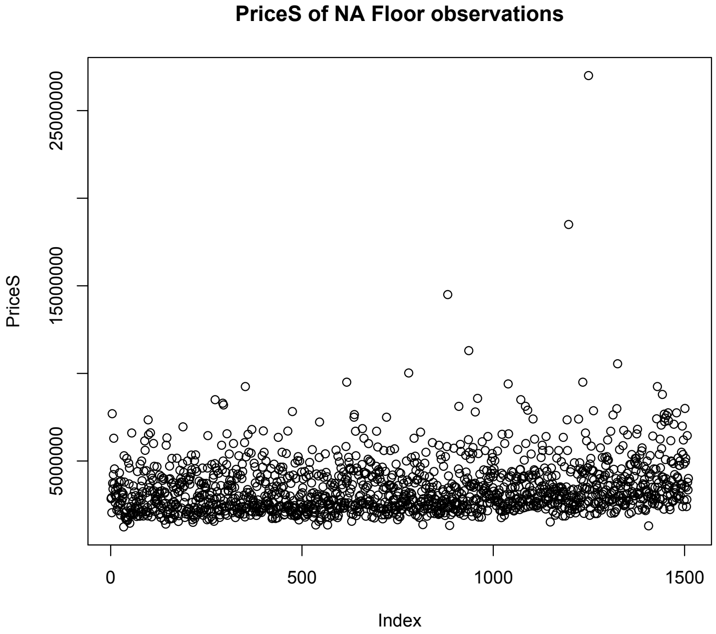

Figure 14: Plot of PriceS of not available Floor observations

Coefficient	Value
(Intercept)	-11999.45254744
PriceArea	0.04071328
AreaRoom	-2.46135341
FloorOne	-7.48077651
FloorTwo	74.89539942
FloorThree	50.72556528
FloorFour	90.33800342
FloorHigh	157.68956727
Area	63.13713095
Age	3.82017464
DateL	0.58580373
Rent	-0.06247285
SeasonWinter	12.84732098
SeasonSpring	47.33467589
SeasonSummer	5.90852202

Table 4: Model 1 Coefficients

Coefficient	Value
(Intercept)	1.17854187817
PriceArea	0.00001207461
logArea	0.94084167204
AreaRoom	-0.00154916996
Rent	-0.00001059427
FloorOne	0.00734908697
FloorTwo	0.02594192111
FloorThree	0.01913642331
FloorFour	0.03361253183
FloorHigh	0.03741263613
Age	0.00057175970
SeasonWinter	0.00609386304
SeasonSpring	0.01075518973
SeasonSummer	0.00562504232
DateL	0.00015546334

Table 5: Log Model 1 Coefficients

Coefficient	Value
(Intercept)	-713.601621300
PriceArea	0.002159004
AreaRoom	-0.293877269
BrokerMedium	-1.588994389
BrokerLarge	-2.331498234
Area	3.832812070
Rent	-0.003498850
FloorOne	1.793863325
FloorTwo	7.104319149
FloorThree	6.138600209
FloorFour	8.219738166
FloorHigh	12.256792962
Age	0.238543738
SeasonWinter	0.769420028
SeasonSpring	3.014175931
SeasonSummer	0.140295956
DateL	0.043840509

Table 6: Box Cox Model 1 Coefficients