An analysis of how variables and home styling affect housing prices

JENNY CHANG
ANDREAS VALDMAA
An analysis of how variables and home styling affect housing prices

JENNY CHANG
ANDREAS VALDMAA

Degree Projects in Applied Mathematics and Industrial Economics
Degree Programme in Industrial Engineering and Management
KTH Royal Institute of Technology year 2018
Supervisor at Widerlöw & Co: Tomas Widerlöw
Supervisors at KTH: Daniel Berglund, Hans Lõof
Examiner at KTH: Henrik Hult
Abstract

Based on the growing interest for home styling and earlier psychological scientific evidence, this study examines how home styling and other variables affect the final price of condominiums in Uppsala. Using multiple linear regression and different statistics, seven different models are analyzed in order to determine whether or not home styling is an influencing factor. To obtain a reliable result, nine other variables such as starting price, living area and floor level etc. are included in the initial model. In addition, these models are investigated statistically to determine if near linear dependence among the regressor variables exists or not. The results show that home styling have a positive impact on the final price of a condominium. The different analytical methods do not always agree, but if looking at the regression result and confidence interval it is obvious that home styling can help increase the final price. Using variable selection, home styling is only included in the model when allowing seven or more variables. The results and analysis from this report is not enough to determine exactly how much home styling affects the final price; since home styling is converted to a dummy variable in the study. The conclusion is that there is a correlation between the response, final price, and the regressor variable, home styling.
Sammanfattning

Acknowledgments

First of all, we, Andreas Valdmaa and Jenny Cheng would like to give special thanks to Tomas Widerlöw at Widerlöw & Co for providing us with data and making this thesis possible. Furthermore we would like to thank Daniel Berglund, at the Institute for Mathematical Statistics, for taking the time to help us and for providing us with valuable information.
# Contents

Abstract .......................... 1

Sammanfattning ..................... 2

Foreword ........................... 3

1 Introduction ......................... 1
   1.1 Background .......................... 1
      1.1.1 Background home styling ............. 1
      1.1.2 Background Economy .................. 3
   1.2 Purpose and Aim ........................ 5
   1.3 Demarcation .......................... 6
      1.3.1 Geographic Limitations ............... 6
      1.3.2 Multicollinearity .................... 6
      1.3.3 Dataset ............................ 6

2 Mathematical Theory and Statistics 8
   2.1 Multiple Linear Regression ............... 8
      2.1.1 Ordinary Least Squares Estimation .... 9
      2.1.2 Necessary Assumptions ................ 9
   2.2 Model Validation and Variable Selection .... 10
      2.2.1 Variables ........................... 10
      2.2.2 Residuals ............................. 10
      2.2.3 Cook’s Distance ....................... 12
      2.2.4 Coefficient of Determination, $R^2$ and Adjusted $R^2$ 12
      2.2.5 Multicollinearity ..................... 13
      2.2.6 AIC and BIC .......................... 14
      2.2.7 Mallow’s CP .......................... 15
      2.2.8 Test-statistics, F-test, t-test, p-value .... 15
      2.2.9 Confidence Intervals CI ............... 16

3 Method ............................ 18
   3.1 Data Collection ....................... 18
   3.2 Variable Description .................... 19
      3.2.1 Dependent variable .................... 19
      3.2.2 Independent variables ................ 19
   3.3 Initial Model ......................... 21
   3.4 Model Validation ...................... 22
   3.5 Variable Selection ..................... 26
4 Results
   4.1 Final Models ........................................ 32

5 Discussion and Conclusion 35
   5.1 Discussion of the Results .......................... 35

6 Conclusion 37
1 Introduction

1.1 Background

1.1.1 Background home styling

"Now the buyer’s market is better than ever before", says the headline of "Veckans affärer"\(^1\). The prices of condominiums have increased significantly during 2017 and are expected to flatten out during 2018.\(^2\)

The housing market is and has been the headline of many articles and news reports. For a long time we have been warned about housing bubbles and an overheated housing market. These articles form the basis for the decision-making of many individuals and get a tremendous role in today’s housing market. A common idea is that the housing prices depend on different conceptual factors such as mortgage rate, economy and area. Thus one may wonder if there are any factors that one can affect as a seller and what actually determines the housing prices.

A featured topic is home styling. Home styling, can by simple means be defined as restyling an apartment before a sale in order to increase the final housing value. It can therefore be interesting to see if it is possible that the same apartment with a different styling can end up with a different final price. According to the American design psychologist, Sally Augustin "home buyers are emotionally influenced, both consciously and unconsciously, by the environment on a housing display, which in turn has an affect on purchasing ability."\(^3\)

From a design psychological scientific perspective a buyer is always looking for a home that they can enjoy and feel safe in. When people buy a home, they try to imagine how it would be like living there in the future and therefore the purchase decision becomes very emotional. We all need a home that suits our own lifestyle and our life goals therefore buying a home has emotional similarities to finding a partner.\(^4\)

As humans we desire to mimic nature when styling a living space and the

\(^1\)Direkt, Veckans affärer, Mäklarna: nu är det köparens marknad mer än någonsin förut, 2017-12-20, (hämtad 2018-03-22)

\(^2\)Direkt, Veckans affärer, Färsk siffer: Så mycket (!) rusade bostadspriserna under 2017, 2018-01-09, (hämtad 2018-03-22)

\(^3\)Augustin S, Homestyling, vetenskaplig homestyling, (hämtad 2018-03-22)

\(^4\)Ibid.
feeling of biophilia is a biological need we have. While with the help of home styling it is possible to create a more natural surrounding indoors. There are four important principals to consider when styling a biophilic home:

- **Control**
  Buyers often like to feel a sense of control when buying an apartment. It is of great importance to have zone divided light switches, openable windows and a balcony so buyers can adjust the temperature and humidity. It is often preferable that the object has thermostats so the temperature can be regulated separately in each room. This provides the immediate feeling of control.

- **Territories**
  To live in an accommodation for over 3 years, clear territories are necessary. In this context territories are areas that the resident can control and have an overview of. A territory does not necessarily need to be distinguished by walls and doors. The division can be accomplished using a piece of furniture that is more eye-catching than the rest. In order to increase well-being it is also important to clarify territory for each individual family member, and if there is a territory for guests or larger groups this should also be clarified.

- **Safety**
  For a comfortable home the interior needs to give a sense of security. To achieve this, one must highlight places of safety, and by doing this it makes the home more attractive to prospective buyers. For example, to attain safety the chairs in the room should be facing the door’s opening so that no one can sneak behind. This way the buyers realize that they will be able to design similar ”safe” spaces once they have moved in.

- **Visual complexity**
  Humans have always enjoyed areas with moderate visual complexity. We should not forget that we interpret our surroundings through our senses. By adhering to controlled order, limited color palettes, small-scale patterns and appealing smells, the visual complexity is maintained at a moderate level. This way the current mood is raised and the stress level is significantly lowered.

---

5Manninen M, Biofilisk inredningsdesign ökar välmåendet, 2017-11-05, (hämtad 2018-04-10)
From a more practical point of view home styling can easily be described as adding furniture and accessories to create a more appealing environment. The interior details must be carefully and thoroughly chosen. Every room should only expose a few sentimental details such as photos, and the interior should be customized by area, feel and purpose.  

1.1.2 Background Economy

With the help of home styling it is possible to increase the value of an object and by doing this the seller’s profit and the broker’s fee increases. Home styling has grown in the Swedish market during the last years and many people see styling as a safe investment to increase the value of an object. A survey by Sifo shows that over 50 percent of the Swedish population believes that that home styling can increase the final price by 10 percent.  

Styling is said to have a positive impact on housing prices and helps increase them, which in turn can affect the economy. Around 85 percent of the household lending consists of mortgages and almost half of the lending to companies goes to commercial real estate in some form, including tenant-owner associations. Therefore the developments in the real estate market have a major impact on the bank’s financial position and the economy.  

Increasing housing prices increases household’s indebtedness, which can lead to exposed risks to housing bubbles and housing shortages. As housing prices rise, consumers are increasingly indebted and many households may have trouble paying their loans and interest rates when interest rate increases. In order to increase the bank’s resistance shaft, the bank can control interest rate and amortization with monetary policy. Today, banks have chosen to follow FI’s proposal and raise the amortization requirement, which means that all new homeowners who borrow more

---

6Augustin S, Home styling, vetenskaplig home styling, (hämntad 2018-04-11)
7Ibid.
8Karlsson M, Dalabygden, Enkel home styling kan påverka bostadspriset, 2018-02-27, (hämntad 2018-03-21)
9Thedéen E generaldirektör FI, Realtid, FI: Höga bostadspriser skapar höga risker, 2017-11-29, (hämntad 2018-04-12)
10Wikipedia, Penningpolitik, 2018-02-16, (hämntad 2018-04-12)
11Thedéen E generaldirektör FI, Realtid, FI: Höga bostadspriser skapar höga risker, 2017-11-29, (hämntad 2018-04-12)
than four and a half times the household’s gross income must amortise an additional one percentage point on the entire mortgage loan in addition to current requirements.\footnote{Nordea, Nytt amorteringskrav från den 1 mars 2018- så påverkar det dig (hämtad 2018-04-14)}

However, it is very difficult to prove that the styling has a significant impact on the final price. Fairly so, styling has a great deal of power in different living areas. Capital buyers who buy in more exclusive areas can certainly afford to pay more or less depending on whether the apartment gives a better or worse impression. However, in areas where people are looking for housing due to housing shortages, it may be difficult to justify a higher price on the basis of better styling, as the decision is usually based on the lowest price. As earlier stated, there is already scientific evidence that the styling has an impact on housing prices. Our task in this study is to show further mathematical and scientific links between home styling and final prices.

In general, the housing prices mostly depend on supply and demand according to economics theories.\footnote{Krugman P Wells R, ECONOMICS, W.H.Freeman Co Ltd, 4th ed 2015} From earlier reports and research it is also known that other factors such as, unemployment, income and mortgage rate also have an impact on the final price of a house. \footnote{Svensk fastighets förmedling, Vad påverkar bostadspriserna?, (Hämtad 2018-05-03)}
1.2 Purpose and Aim

Our study aims to see how different prognostic factors within condominiums change over time. Over the past years home styling has become a more common action in the housing market. This option is offered by both brokerage firms and professional styling companies. Among real estate agencies it is thought to be a factor that can help increase the final price. Hence the primary purpose will be to investigate whether or not home styling is a crucial factor for the final price of a condominium. From a business point of view, this is highly relevant and can be significant to both broker firms and private customers. In the long run this may even have an impact on the economy.

In this study we aim to find answers to the questions below:

- Is there a linear relationship between home styling and the response, final price?
- Is it scientifically recommended to invest in home styling based on our results?
- To what extent does home styling influence the final price?
- How can home styling contribute to the society and economy?
1.3 Demarcation

1.3.1 Geographic Limitations

The sample of data points in this study are collected from the city Uppsala in Sweden. By limiting this study to one city it is easier to control the data and to analyze the results. However by only investigating one city, some limitations may appear. It is hard to apply the results from this study to other cities, since the results from this study is only based on Uppsala, and may not be consistent with other cities. Since home styling costs extra money in some extent, it is to be considered as a luxury investment. It is therefore also more accessible to the upper middle class and a tool used in greater extent in richer cities and areas. Uppsala is the fourth biggest city in Sweden\textsuperscript{15}, but it still differs very much from the capital city, Stockholm. There are more wealthy people, more expensive houses and more interior agencies in a big city like Stockholm. This way it only seems natural that home styling is more standardized and widespread in Stockholm.

1.3.2 Multicollinearity

Some of the variables in the model are probably highly correlated with each other. For example, area and rooms are probably highly correlated, which impairs the reliability of the regression. Another limitation could be that brokers may automatically set a higher starting price if the object has been styled. This way it is hard to show if home styling has an impact or not.

1.3.3 Dataset

Two different data sets were provided, thus these had to be merged into one. Originally, the data set contained 10700 data points, collected between 2007 and 2017, however only four of these years contained information about styling. Therefore the data set was reduced to only 2161 data points, collected between 2015-2017. When merging the data set some points were lost, since they did not match. This way the data set

\textsuperscript{15}Världens häftigaste, Byggnadsverk, Sveriges 15 största städer, 2018-03-28, (hämtad 2018-05-06)
ended up containing only 166 styled objects. The model adequacy becomes more accurate when more points are included. Since we have few data points of home styling the results become uncertain and difficult to interpret. In the data set there are only condominiums and other objects have been deleted.
2 Mathematical Theory and Statistics

All the theory under this headline is based on the literature "Introduction to Linear Regression Analysis".\textsuperscript{16} unless otherwise stated.

The purpose of this thesis is to investigate and to model the relationship between several different factors and the final price of a condominium. This is achievable with one of the most widely used techniques for analyzing multiple factor data, namely regression analysis. It’s application originates from the logical process of using an equation to express the correlation between a response variable and a set of regressors, also called prediction variables. Linear regression describes how the response variable depends on the prediction variables.

2.1 Multiple Linear Regression

The multiple linear regression is used when there are more than two measurable variables. In this study there is one dependent variable and several independent variables, hence multiple linear regression is applied. The definition of the multiple linear regression, given n observations, is

\begin{equation}
y_i = \sum_{i=1}^{n} x_{ij} \beta_j + e_i \quad i = 1, 2, ..., n
\end{equation}

where \(y_i\) is the dependent variable, \(x_{ij}\) is the value of the independent variables and \(\beta_j\) is the unknown parameters which will be estimated by the data. The random error components are given by \(e_i\) and are assumed to be uncorrelated and all the normality assumptions are assumed to hold. The model can be represented in a more compact matrix form, in matrix notations the model is given by:

\begin{equation}
y = X\beta + \epsilon
\end{equation}

where

\[ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \] (3)

\[ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \] (4)

### 2.1.1 Ordinary Least Squares Estimation

A simple way to estimate the unknown regression coefficients, \( \beta_i \), is using the Ordinary Least Squares estimation (OLS). In the method the optimal estimate is obtained by minimizing the sum of the squared residuals i.e. minimizing the square distance between the observed value \( y_i \) and the fitted value \( \hat{y}_i \). The best linear unbiased estimator (BLUE), that is, the estimator with the minimum variance among those that are unbiased and linear combinations of \( y_i \) can only arise under certain conditions by the Gauss Markov theorem. Given that the explanatory variables are linearly independent and the inverse of the correlation matrix \( (X'X)^{-1} \) exists, the vector of the least-square estimates \( \hat{\beta} \) present the best estimate possible for the relationship between the dependent variable and the independent variables. \( \hat{\beta} \) is mathematically estimated by minimizing:

\[ S(\beta) = (y - X\beta)'(y - X\beta) \] (5)

\[ \frac{\partial S}{\partial \beta} \bigg|_{\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0 \] (6)

\[ \hat{\beta} = (X'X)^{-1}X'y \] (7)

### 2.1.2 Necessary Assumptions

For a multiple linear regression model there are five necessary assumptions needed.
1. **Linear Relationship**  
The linear functional form is correct, i.e. the dependent variable $y$ and the independent variables $X_i$ are linear.

2. **Strict exogeneity**  
The expected value of the error term in the regression should have conditional mean zero, i.e. is unbiased.

3. **No Multicollinarity**  
Multicollinarity does not exist and the regressors in $X$ must all be linearly independent.

4. **Homoscedasticity**  
The error term has the same variance in each observation.

5. **No auto-correlation**  
The errors are uncorrelated between the observations.

### 2.2 Model Validation and Variable Selection

#### 2.2.1 Variables

The explanatory variables can be characterized by standard or dummy variables. The standard version of a variable is that it is defined by a continuous numerical value, and a dummy variable is classified by being an indicator or a categorical variable. This means that the dummy variable is only taking the values of one or zero depending on if the information should be included or not in the model.

The linear relationship has a slightly different meaning between a continuous variable and a categorical variable which implies that linearity is always met, however by examining the coefficient $\beta$ it can be concluded if it is significant or not.

#### 2.2.2 Residuals

The classical definition of residuals are:

$$ e_i = y_i - \hat{y}_i, i = 1, ..., n $$  \hspace{1cm} (8)

$y_i$ are the observations and $\hat{y}_i$ are the corresponding fitted values. The residual can be described as the deviation between the data and the fit.
We use residuals to discover different types of model inadequacies. The residuals have mean zero and the approximate average variance is estimated by:

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

(9)

It may be handy to scale the residual in order to easier identify outliers and extreme values later.

Three common scalings are:

- Standardized residuals
- Studentized residuals
- PRESS residuals

**Standardized residuals** - If scaling the residual with the approximate average variance \(MS_{Res}\) the result will be the standardized residual:

\[
d_i = \frac{e_i}{\sqrt{MS_{Res}}} , i = 1, 2, ... n
\]

(10)

The standardized residuals have zero mean and approximately unit variance. A large standardized residual, \(d_i > 3\), potentially indicates an outlier. In this case we need to look into these specific points further.

**Studentized residuals** - By scaling the residual, \(e_i\) with the exact standard deviation of the \(i\)th residual, instead of the approximate standard deviation \(MS_{Res}\), the scaling can be improved. The studentized residual may be viewed as:

\[
r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}} , i = 1, 2, ... n
\]

(11)

The studentized residuals have constant unit variance regardless of the location of \(x_i\) when the model is correct. A point with a large residual and a large \(h_{ii}\) is potentially high influential.

**PRESS residual** - The prediction error also so called PRESS residual can be described as:

\[
e_i = y_i - \hat{y}_{(i)}
\]

(12)
where \( \hat{y}_{(i)} \) is the fitted value of the \( i \)th response based on all observations except the \( i \)th one. To determine which points that are to be considered influential the prediction error calculation is repeated for each observation \( i=1,2,...,n \). Since the \( i \)th observation is deleted then \( \hat{y}_{(i)} \) cannot be influenced by that specific observation. This way the resulting residual will likely indicate the presence of an outlier. Consequently, large PRESS residuals are generally high influence points.

### 2.2.3 Cook’s Distance

An observation point that considerably departs from the dataset is called an outlier. A potential y space outlier can be recognized by residuals that are three or four standard deviations from the mean in absolute value. Outliers in a model should be carefully investigated, depending on their location in x space they can have severe effects and that is why you want to decide whether the abnormality of the outlier make sense or not.

Outliers can be classified either as a leverage- or influence-point. A leverage point can be defined as an outlier in x space but with residuals that are practically equal to the mean. It means that a small amount of change of it’s position causes a large change in the model behavior. A influence point in the other hand has both moderately unusual coordinates in x- and y-space.

There is several techniques to detect outliers, Cooks distance is one of them. The Cooks distance is a measure of the squared distance between least squares based on all \( n \) points \( \hat{\beta} \) and the estimate obtained by deleting the \( i \)th point, say \( \hat{\beta}_i \). The formula for Cooks distance is:

\[
D_i = (M_i, c) = (\hat{\beta}_i - \hat{\beta})' M (\hat{\beta}_i - \hat{\beta}) + c
\]

A general rule of thumb is that a point with a Cook’s \( D_i \) of more than 4 times the mean could be a possible outlier. Others say that values of \( D_i \) larger than 1 indicates an influential point and that values above 0.5 should be investigated.

### 2.2.4 Coefficient of Determination, \( R^2 \) and Adjusted \( R^2 \)

\( R^2 \) also called the coefficient of determination, is a statistical measure of the goodness of fit, i.e. the correlation between the dependent variable and the covariates. The \( R^2 \)-coefficient indicates the proportion of
variation of $ly$ that can be explained by the independent variables. $R^2$ is defined as

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$  \hspace{1cm} (13)$$

where

$$SS_T = SS_R + SS_{Res}$$  \hspace{1cm} (14)$$

$$SS_R = \hat{\beta}'X'y - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2$$  \hspace{1cm} (15)$$

$$SS_{Res} = y'y - \hat{\beta}'X'y$$  \hspace{1cm} (16)$$

$SS_T$ measures the total variability in the observations, i.e. the corrected sum of squares of the observations.

$SS_R$ is the regression sum of squares, $SS_T$ is the corrected sum of squares of the observations and $SS_{Res}$ is the residual sum of squares. Evidently, the fit of the model will be better if the residual sum of square is minimized. $R^2$ is generally said to be the proportion of variation in the response variable explained by the explanatory variables. Since $0 < SS_{Res} < SS_T$, it implicates $0 < R^2 < 1$, where a value close to 1 indicates that most of the variability in the response variable can be explained by the model.

In general, the higher the R-squared, the better the model fits your data.

\textbf{2.2.5 Multicollinearity}

Multicollinearity or near-linear dependence among the regressors affect the usefulness of a regression model. Multicollinearity can be detected by using the multicollinearity diagnostics, variance inflation factors VIFs, which are the main diagonal elements of the inverse of the $X'X$ matrix in correlation form, $(W'W)^{-1}$.

$$VIF_j = \frac{1}{1 - R_j^2}$$  \hspace{1cm} (17)$$
where $R^2_j$ denotes the coefficient of determination obtained from regressing $x_j$ on the other regressors.

Regression models based on the least square method provide poor prediction variables and the values of the estimates are also very sensitive to the data sample when strong multicollinearity is present. VIFs larger than 10 imply serious problems with strong multicollinearity. If $VIF_1 = VIF_2 = 1$, there is no linear relationship between the regressor 1 and 2, and they are said to be orthogonal.

**There are four primary sources of multicollinearity:**

1. The data collection method employed, meaning that samples are only subspaces of the region of the regressors defined
2. Constraints on the model or in the population, i.e. physical constraint on the sample can cause multicollinearity
3. Model specification, for example when adding polynomial terms to a regression model it causes ill-conditioning if the range of $x$ is small
4. An overdefined model, meaning that there are more regressor variables than observations

### 2.2.6 AIC and BIC

Akaike Information Criterion (AIC) estimates the quality of a model relative to each of the other models. AIC is useful measure when performing variable selection. The AIC criterion is defined as:

$$AIC = 2k + nln|\epsilon|^2$$  \hspace{1cm} (18)

Another criterion that is often used is Bayesian Information Criterion (BIC):

$$BIC = kln(n) + nln\left(\frac{SS_{Res}}{n}\right)$$  \hspace{1cm} (19)

where $n$ is the number of observations and $k$ is the number of predictor variables.

The preferred model is the one with the minimum AIC and BIC values.
2.2.7 Mallow’s CP

Mallow’s CP is another criterion that has a means when executing model selection. Mallows’s CP has been shown to be equivalent to AIC in some special cases of Gaussian linear regression.

The criterion is related to the mean square error of the fitted value, as follows:

\[ E[\hat{y}_i - E(y_i)]^2 = [E(y_i) - E(\hat{y}_i)]^2 + \text{Var}(\hat{y}_i) \]  

(20)

\( E[y_i] \) is the expected response from the true regression equation and \( E[\hat{y}_i] \) is the expected response from the subset model, therefore \( E(y_i) - E(\hat{y}_i) \) becomes the bias at the i:th data point.

Small value of \( Cp \) is desirable and means that the model is relatively precise.

2.2.8 Test-statistics, F-test, t-test, p-value

F-test

To determine if there is a linear relationship between the response \( y \) and any of the regressors \( x_1, x_2, ..., x_k \) we test the significance of the regression by doing a global test of model adequacy.

If at least one of the regression coefficient is statistically significant, then the null hypothesis can be rejected:

\[ H_0 : \beta_1 = ... = \beta_k = 0 \]  

(21)

To reject the null hypothesis the F statistic need to be computed:

\[ F_0 = \frac{SSR/k}{SS_{Res}/(n-k-1)} = \frac{MSR}{MSE} \]  

(22)

\( F_0 \) follows a \( F_{k,n-k-1} \) distribution and if \( F_0 > F_{a,k,n-k-1} \) then the null hypothesis should be rejected.

t-test
If the F-test shows that at least one of the regressors are significant for the regression model, we need to know which one. When adding a variable to the model the sum of squares increases and the residual sum of squares decreases. This also causes a increase in the variance of the fitted value $\hat{y}_i$. Adding an irrelevant regressor can affect the residual mean square negatively and decrease the usefulness of the model. This form the basis for why we need to check the significance of each individual coefficient.

When testing the significance of an individual regression coefficient the hypothesis will be:

$$H_0 : \beta_j = 0, H_1 : \beta_j \neq 0$$ (23)

the t-statistic for this hypothesis is:

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, \quad \hat{\beta}_j \text{ se}(\hat{\beta}_j)$$ (24)

Where $C_{jj}$ is the diagonal element of $(X'X)^{-1}$ corresponding to $\hat{\beta}_j$. If $|t_0| = t_{\alpha/2,n-k-1}$ then the null hypothesis should be rejected.

**P-value**

From the statistics the P-value can be calculated. P-value tells if the coefficient is significant relative the significance level $\alpha$.

$$p = Pr(X \geq F), X \in F_{k,n-k-1}$$ (25)

where X is a random variable. This tells the probability of X being greater than the F-value. If the p-value is smaller than the significance level $\alpha$ (often 0.05), then the null hypothesis should be rejected.

### 2.2.9 Confidence Intervals CI

Based on the statistics a 100$(1 - \alpha)$ confidence interval for each regression coefficient can be calculated.

If all the normality assumption hold then the CI is defined as below:
\[
\hat{\beta}_j - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \tag{26}
\]

where

\[
\sqrt{\hat{\sigma}^2 C_{jj}} = Se(\hat{\beta}_j) \tag{27}
\]
3 Method

This section will clarify our principal strategy and our main methods to approach the different subjects of this thesis. Because of the structure of our data set and the few data points from home styling we will obtain two different sets, which will stand as base for the fitted models that we will be evaluating. The strategy for variable selection and model building that has been applied on the models is represented by following steps\textsuperscript{17}:

1. Fit the largest model possible to the data.
2. Perform a thorough analysis of this model.
3. Determine if a transformation of the response or of some of the regressors are necessary.
4. Determine if all possible regressions are feasible.
5. Compare and contrast the best models recommended by each criterion.
6. Perform a thorough analysis of the “best” models.

3.1 Data Collection

The data set we examined was provided by Widerlöv & Co which in turn has been collected from their customer database. Throughout we acquired a deeper understanding about the housing market and the pricing strategies by communicating with Widerlöv & Co and through individual research. All together this led to a more thorough knowledge about the data set variables. The original data was split in two separate files, one with sale statistics over the period 2007-2017 and one with styling statistics over the period 2014-2018.

To be able to accomplish a regression analysis a complete data set is needed, therefore these two sets where merged into one. Because of the different structure of the explanatory variables, dummy variables were created for floor level, location, home styling and construction year. Due to insufficient data (including missing and squint values) for some of the chosen explanatory variables the set was reduced from \(M\) numbers of observations to \(N\) numbers of observations. These procedures were required to reduce and group the data in order to make assumptions.

\textsuperscript{17}Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, WILEY, Introduction to Linear Regression Analysis, 5th ed. 2012, section 10.3
3.2 Variable Description

3.2.1 Dependent variable

Final price $SEK/m^2$

The final price for the condominiums is set as the dependent variable. This is obvious when the intention with this thesis is to consolidate the fact that home styling has an impact on the housing market. The relationship between the final price and home styling can possibly prove that home styling does have a profitable effect on the price. Therefore, the final price will be our output i.e. response variable.

3.2.2 Independent variables

All the variables below are independent variables and also so called prediction variable. They should not depend on other factors. The prediction variables are given in the data received from Widerlöv & Co

1. Living Area $m^2$

Area defines the total living area of the object in square meter. The final price depends on the living area. The theory says that larger the surface the higher the final price, but at the same time the price per square meter will be less.

2. Starting Price $SEK/m^2$

The starting price is the price when the object is first placed on the market. It feels logical that a house with a higher starting price is also sold at a higher final price. However, this is not always true, since the starting price depends on the brokers evaluation technique.

3. Number of Rooms $R$

The number of rooms in each condominiums are given in integers such as 1, 2, 3 etc.
4. Floor Level

Since apartments on ground level tend to sell for lower prices it seems reasonable to set the floor level as a dummy variable i.e. (0) or (1). 1 represents that an apartment lies on the first floor or below and 0 means that the apartment lies above the first floor.

5. Rental Fee $SEK$

The rental fee is a fee payed to the housing association, which covers the operating cost of the entire residential building. For example the fee goes to reparations, wash house, garbage collection and road maintenance. This free often varies with the construction year and the housing association’s economy. Therefore it is very likely that we will detect multicollinearity between the rental fee and construction year, which in that case need to be taken care of.

6. Location

The location of an accommodation is often highly influential on the final price. Here we have encountered some problems since there are countless many different living areas in Uppsala. We chose to set the location as a dummy variable where 1 represents that the object is located down town and 0 shows that the object is located outside the inner city.

7. Construction year

In the data, the constructions years lies within the range 1825-2017. We have grouped the variable into five smaller intervals depending on the construction year:

- Dummy variable A takes value 1 if it occurs within the interval 1800-1899 and 0 if not
- Dummy variable B takes value 1 if it occurs within the intervall 1900-1949 and 0 if not
- Dummy variable C takes value 1 if it occurs within the interval 1950-1999 and 0 if not
• If all dummy variables are 0 the object belongs to the interval 2000-2017

Since most people values the construction year it is of great importance to create reasonable intervals. The division is based on the fact that people often value turn of century buildings and new buildings very high.

8. Home styling

The main purpose of this project is to show that home styling has a significant impact on the final price. There are many different types of home styling and interior techniques, but for the sake of simplicity one dummy variable is created. All objects that have been styled receive the value 1 and non-styling objects the value 0.

3.3 Initial Model

<table>
<thead>
<tr>
<th>Variables</th>
<th>Estimates</th>
<th>Std.Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5998.0000</td>
<td>51.5200</td>
<td>11.6420</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>Starting Price</td>
<td>0.9722</td>
<td>0.0087</td>
<td>112.3680</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>Number of Rooms</td>
<td>341.4000</td>
<td>180.0000</td>
<td>1.8970</td>
<td>0.0580  .</td>
</tr>
<tr>
<td>Area</td>
<td>-34.5900</td>
<td>9.5970</td>
<td>-3.6040</td>
<td>0.0032  ***</td>
</tr>
<tr>
<td>Floor Level</td>
<td>423.0000</td>
<td>147.6000</td>
<td>2.8650</td>
<td>0.0042  **</td>
</tr>
<tr>
<td>Rental Fee</td>
<td>-0.2608</td>
<td>0.1223</td>
<td>-2.1320</td>
<td>0.0331  *</td>
</tr>
<tr>
<td>Location</td>
<td>1117.0000</td>
<td>275.8000</td>
<td>4.0520</td>
<td>0.0001  ***</td>
</tr>
<tr>
<td>Construction 1800-1899</td>
<td>-21.4900</td>
<td>688.0000</td>
<td>-0.0310</td>
<td>0.9751  .</td>
</tr>
<tr>
<td>Construction 1900-1949</td>
<td>902.5000</td>
<td>165.5000</td>
<td>5.4540</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>Construction 2000-2017</td>
<td>-1228.0000</td>
<td>182.7000</td>
<td>-6.7180</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>home styling</td>
<td>721.7000</td>
<td>329.4000</td>
<td>2.1910</td>
<td>0.0286  *</td>
</tr>
</tbody>
</table>

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  

Table 1: Summary Table

With all the variables above, the initial model to describe the final price per $m^2$ is represented by:

\[ Y_{FinalPrice_i} = \beta_0 + \beta_1 x_{StartPrice_i} + \beta_2 x_{LivArea_i} + \beta_3 x_{Rooms_i} + \beta_4 x_{Floor_i} + \beta_5 x_{Fee_i} + \beta_6 x_{Location_i} + \beta_7 x_{1800-1899} + \beta_8 x_{1900-1949} + \beta_9 x_{2000-20017} + \beta_{10} x_{homestyling_i} + \epsilon_i, \]

where \( i = 1, \ldots, n, n = 2161 \)

The model in a more compact matrix form is described by:
\[ Y_{\text{FinalPrice}} = \beta X + \epsilon \] (28)

where

\[ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \] (29)

\[ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \] (30)

where \( n = 2161, k = 11 \)

3.4 Model Validation

Linearity and Homoscedasticity Assumptions

The assumption of linear relationship must be coherent with the model and this can be accomplished by studying a scatter plot. The scatterplot shows the residuals on the y axis and the fitted values on the x axis of the model. By studying the plot it can be seen that the data seems to be well modelled by a linear relationship, and that the points appear to be randomly spread out around the line, with no distinct non-linear trends.
From the Scale-location plot where the square root of the standardized residuals is plotted on the y axis and the fitted values on the x axis we see that the spread of the residuals are equally random distributed. It is almost possible to identify a funnel pattern, but still we seem to have an even spread of residuals along the ranges of predictors. Since there is a horizontal line with equally spread points, the assumption of equal variance, homoscedasticity, seems to hold.
Normality Assumptions

A good approach to decide whether a possible transformation is needed or not is an investigation of the residual plot. The residual plot is literally the residuals $e_i$ plotted against the fitted values $\hat{y}$. The plots below show two different transformations, where no linear relationship is to prefer. By studying these plots one can conclude that there is definitely no improvement of linearity.

![Figure 3: Log Transformation and Squared Transformation](image)

By exploring the Q-Q plot conclusions concerning normally distributed residuals can be made. From the plot it is easy to see that non-transformed model provides a Q-Q plot where the line is relatively straight and diagonal. Thus the plot implies that the response variable is positive and varies largely by size. By this it can be concluded that normality assumptions is verified.

![Figure 4: The Q-Q plot](image)
When investigating the outliers we can see that these points are different from the mean but do not exceed the value of the critical distance of 0.5, though this do not imply any abnormality. The values can be motivated and can therefore not be excluded from the dataset.
3.5 Variable Selection

Multicollinearity

To determine if multicollinearity is present in the model it was examined through the variance inflation factor VIF. From table 1 multicollinearity was detected by large values on both Living Area and Number of Rooms. All values larger than 10 indicate multicollinearity and therefore Living Area was eliminated from the initial model.

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting Price</td>
<td>2.174047</td>
</tr>
<tr>
<td>Living Area</td>
<td>12.403793</td>
</tr>
<tr>
<td>Number of Rooms</td>
<td>8.150796</td>
</tr>
<tr>
<td>Floor Level</td>
<td>1.019577</td>
</tr>
<tr>
<td>Rental Fee</td>
<td>4.981814</td>
</tr>
<tr>
<td>Location</td>
<td>1.130939</td>
</tr>
<tr>
<td>Construction 1800-1899</td>
<td>1.098131</td>
</tr>
<tr>
<td>Construction 1900-1949</td>
<td>1.461116</td>
</tr>
<tr>
<td>Construction 2000-2017</td>
<td>1.467320</td>
</tr>
<tr>
<td>home styling</td>
<td>1.017824</td>
</tr>
</tbody>
</table>

Table 2: VIF values, Initial Model

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting Price</td>
<td>2.026884</td>
</tr>
<tr>
<td>Number of Rooms</td>
<td>3.199940</td>
</tr>
<tr>
<td>Floor Level</td>
<td>1.015323</td>
</tr>
<tr>
<td>Rental Fee</td>
<td>3.993933</td>
</tr>
<tr>
<td>Location</td>
<td>1.110594</td>
</tr>
<tr>
<td>Construction 1800-1899</td>
<td>1.082251</td>
</tr>
<tr>
<td>Construction 1900-1949</td>
<td>1.459611</td>
</tr>
<tr>
<td>Construction 2000-2017</td>
<td>1.466818</td>
</tr>
<tr>
<td>home styling</td>
<td>1.017618</td>
</tr>
</tbody>
</table>

Table 3: VIF values, Initial Model minus Living Area

When Living Area was deleted from the initial model, the remaining variables were no longer showing sign of multicollinearity.
Tests and Statistics

<table>
<thead>
<tr>
<th>Variables</th>
<th>Estimates</th>
<th>Std.Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5998.0000</td>
<td>51.5200</td>
<td>11.6420</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>Starting Price</td>
<td>0.9722</td>
<td>0.0087</td>
<td>112.3680</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>Number of Rooms</td>
<td>341.4000</td>
<td>180.0000</td>
<td>1.8970</td>
<td>0.0580  .</td>
</tr>
<tr>
<td>Area</td>
<td>-34.5900</td>
<td>9.5970</td>
<td>-3.6040</td>
<td>0.0032  ***</td>
</tr>
<tr>
<td>Floor Level</td>
<td>423.0000</td>
<td>147.6000</td>
<td>2.8650</td>
<td>0.0042  **</td>
</tr>
<tr>
<td>Rental Fee</td>
<td>-0.2608</td>
<td>0.1223</td>
<td>-2.1320</td>
<td>0.0331  *</td>
</tr>
<tr>
<td>Location</td>
<td>1117.0000</td>
<td>275.8000</td>
<td>4.0520</td>
<td>0.0001  ***</td>
</tr>
<tr>
<td>Construction 1800-1899</td>
<td>-21.4900</td>
<td>688.0000</td>
<td>-0.0310</td>
<td>0.9751  .</td>
</tr>
<tr>
<td>Construction 1900-1949</td>
<td>902.5000</td>
<td>165.5000</td>
<td>5.4540</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>Construction 2000-2017</td>
<td>-1228.0000</td>
<td>182.7000</td>
<td>-6.7180</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>home styling</td>
<td>721.7000</td>
<td>329.4000</td>
<td>2.1910</td>
<td>0.0286  *</td>
</tr>
</tbody>
</table>

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

Table 4: Summary Table

Looking at the table above it can be concluded that the regressor area has “wrong” sign, and this is most likely because of the existence of multicollinearity. Area and room has a strong linear dependence, which also seems logical since the more rooms the bigger an apartment tend to be.

The p-value shows that styling has significant code 0.01, which means that there are a significant linear relationship between the respond $y_{\text{FinalPrice}}$ and regressor $x_{\text{Styling}}$.

Since we know from the VIF’s that area and number of rooms are high correlated, area will be reduced from the model. If only looking at the p-value we should also reduce construction year A from the model.

All Possible Regression

For optimal variable selection the all possible regression procedure was performed. The all possible regression method fits all the different conceivable regression equations involving one variable, two variables and so on. These equations can then be evaluated by diverse criterion’s such as Mallow’s CP, $R^2$, adjusted $R^2$ and BIC. By Table 4, it can be adressed that the best model that includes home styling is when the model contains seven variables i.e the initial model except Living Area, Number of Rooms and Construction 1800-1899. This also verify our conclusions from previous methods.
<table>
<thead>
<tr>
<th>Number of Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting Price</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Number of Rooms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Floor Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Rental Fee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Construction 1800-1899</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construction 1900-1949</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Construction 2000-2017</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>home styling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Table 5: All Possible Regression

From this results five new models were created and more thoroughly investigated.

Model Update

Based on previous method’s and statistics (VIF, p-value and All possible regression), the model is being reduced to several reduced models that only include small subsets of the original model. After all possible regression we decided to look into 6 different more optimal models depending on how many variables included in the model.

When detecting correlation between area and number of rooms, area was removed from the model, which resulted in reduced model 1. To determine if a model including home styling give better result, home styling is removed from reduced model 1, which gives reduced model 2. Reduced model 1 and reduced model 2 will later on be compared. When looking at the result from all possible regression one can see that construction year A do not have any impact on the model. So construction year A is removed from reduced model 1. For reduced model 4 area, construction year A and home styling is removed. Reduced model 5 is the smallest model including home styling, when conducting all possible regression. Later we saw that also room only had a very small influence on the final price so we decided to remove that variable too, which give us reduced model 6.

From the analysis, one full model and six different reduced models are found:

1. **Full model**
\[ Y_{\text{FinalPrice}} = \beta_0 + \beta_1 x_{\text{StartPrice}} + \beta_2 x_{\text{Rooms}} + \beta_3 x_{\text{LivArea}} + \beta_4 x_{\text{Floor}} + \beta_5 x_{\text{Fee}} + \beta_6 x_{\text{Location}} + \beta_7 x_{1800-1899} + \beta_8 x_{1900-1949} + \beta_9 x_{1950-1999} + \beta_{10} x_{\text{homestyling}} + \epsilon_i, \]

2. Reduced model 1
\[ Y_{\text{FinalPrice}} = \beta_0 + \beta_1 x_{\text{StartPrice}} + \beta_2 x_{\text{Rooms}} + \beta_3 x_{\text{Floor}} + \beta_4 x_{\text{Fee}} + \beta_5 x_{\text{Location}} + \beta_6 x_{1800-1899} + \beta_7 x_{1900-1949} + \beta_8 x_{1950-1999} + \beta_{10} x_{\text{homestyling}} + \epsilon_i, \]

3. Reduced model 2
\[ Y_{\text{FinalPrice}} = \beta_0 + \beta_1 x_{\text{StartPrice}} + \beta_2 x_{\text{Rooms}} + \beta_3 x_{\text{Floor}} + \beta_4 x_{\text{Fee}} + \beta_5 x_{\text{Location}} + \beta_6 x_{1800-1899} + \beta_7 x_{1900-1949} + \beta_8 x_{1950-1999} + \epsilon_i, \]

4. Reduced model 3
\[ Y_{\text{FinalPrice}} = \beta_0 + \beta_1 x_{\text{StartPrice}} + \beta_2 x_{\text{Rooms}} + \beta_3 x_{\text{Floor}} + \beta_4 x_{\text{Fee}} + \beta_7 x_{1900-1949} + \beta_8 x_{1950-1999} + \beta_{10} x_{\text{homestyling}} + \epsilon_i, \]

5. Reduced model 4
\[ Y_{\text{FinalPrice}} = \beta_0 + \beta_1 x_{\text{StartPrice}} + \beta_2 x_{\text{Rooms}} + \beta_3 x_{\text{Floor}} + \beta_4 x_{\text{Fee}} + \beta_7 x_{1900-1949} + \beta_8 x_{1950-1999} + \epsilon_i, \]

6. Reduced model 5
\[ Y_{\text{FinalPrice}} = \beta_0 + \beta_1 x_{\text{StartPrice}} + \beta_4 x_{\text{Fee}} + \beta_5 x_{\text{Location}} + \beta_7 x_{1900-1949} + \beta_8 x_{1950-1999} + \epsilon_i, \]

7. Reduced model 6
\[ Y_{\text{FinalPrice}} = \beta_0 + \beta_1 x_{\text{StartPrice}} + \beta_3 x_{\text{Floor}} + \beta_4 x_{\text{Fee}} + \beta_5 x_{\text{Location}} + \beta_7 x_{1900-1949} + \beta_8 x_{1950-1999} + \beta_{9} x_{\text{homestyling}} + \epsilon_i, \]

For better overview see table below:

<table>
<thead>
<tr>
<th>Variables</th>
<th>Full</th>
<th>Reduced 1</th>
<th>Reduced 2</th>
<th>Reduced 3</th>
<th>Reduced 4</th>
<th>Reduced 5</th>
<th>Reduced 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting Price</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Number of Rooms</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Area</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Floor Level</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rental Fee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construction 1800-1899</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construction 1900-1949</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construction 2000-2017</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home styling</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of Variables</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 6: Model Summary
AIC, BIC and $R^2$

<table>
<thead>
<tr>
<th>Models</th>
<th>$R^2$</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>0.9358</td>
<td>40840.18</td>
<td>40908.32</td>
</tr>
<tr>
<td>Reduced 1</td>
<td>0.9354</td>
<td>40851.2</td>
<td>40913.66</td>
</tr>
<tr>
<td>Reduced 2</td>
<td>0.9353</td>
<td>40853.77</td>
<td>40910.55</td>
</tr>
<tr>
<td>Reduced 3</td>
<td>0.9354</td>
<td>40849.41</td>
<td>40906.2</td>
</tr>
<tr>
<td>Reduced 4</td>
<td>0.9353</td>
<td>40852.07</td>
<td>40903.18</td>
</tr>
<tr>
<td>Reduced 5</td>
<td>0.935</td>
<td>40857.57</td>
<td>40897.32</td>
</tr>
<tr>
<td>Reduced 6</td>
<td>0.9353</td>
<td>40849.53</td>
<td>40900.64</td>
</tr>
</tbody>
</table>

Table 7: Table of AIC, BIC and $R^2$

Comparing reduced model 1 with reduced model 2 we see that the R-square is slightly better for the model with styling (Reduced model 1) than without styling (Reduced Model 2). If looking at AIC and BIC we see that they differ very little, but do not share the same result. Since R-squared and AIC share the same result we have chosen to rely on this to achieve a final result. When comparing reduced model 3 and reduced model 4, the same conclusion can be drawn. It is though concerning that BIC contradicts the other result, so this will be further investigated in the discussion.

Confidential Interval CI

<table>
<thead>
<tr>
<th>Variables</th>
<th>Estimates</th>
<th>Std.Error</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5.408e+03</td>
<td>4.898e+02</td>
<td>4447.2531816</td>
<td>6368.4757661</td>
</tr>
<tr>
<td>Starting Price</td>
<td>9.804e-01</td>
<td>8.378e-03</td>
<td>0.9639261</td>
<td>0.9967841</td>
</tr>
<tr>
<td>Number of Rooms</td>
<td>-1.642e+02</td>
<td>1.131e+02</td>
<td>-386.0242183</td>
<td>57.5701870</td>
</tr>
<tr>
<td>Floor Level</td>
<td>3.887e+02</td>
<td>1.477e+02</td>
<td>98.9528778</td>
<td>678.3976434</td>
</tr>
<tr>
<td>Rental Fee</td>
<td>-4.572e-01</td>
<td>1.098e-01</td>
<td>-0.6725802</td>
<td>-0.2417509</td>
</tr>
<tr>
<td>Location</td>
<td>9.841e+02</td>
<td>2.740e+02</td>
<td>446.7110789</td>
<td>1521.5705035</td>
</tr>
<tr>
<td>Construction 1800-1899</td>
<td>-3.197e+02</td>
<td>6.849e+02</td>
<td>-1662.9105903</td>
<td>1023.5035379</td>
</tr>
<tr>
<td>Construction 1900-1949</td>
<td>9.217e+02</td>
<td>1.658e+02</td>
<td>596.4243185</td>
<td>1246.8964558</td>
</tr>
<tr>
<td>Construction 2000-2017</td>
<td>-1.240e+03</td>
<td>1.832e+02</td>
<td>-1599.0714241</td>
<td>-880.4592200</td>
</tr>
<tr>
<td>home styling</td>
<td>7.048e+02</td>
<td>3.302e+02</td>
<td>57.1590100</td>
<td>1352.3798775</td>
</tr>
</tbody>
</table>

Table 8: Table of Confidence Intervals

The CI for styling is positive and does not include zero, which means that styling has a significant effect on the final price. But we also see
that the CI is very wide and therefore it is hard to say how big influence the variable actually have.
4 Results

In the following chapter the results that were obtained with the methods described in the method chapter are compiled, analyzed and compared with the existing knowledge and theory presented in the frame of the reference chapter. The results will be presented through the chosen statistical tests, graphs and tables.

4.1 Final Models

By applying different types of variable selection methods on the initial model two reduced models could be obtained as a result. The concluding models were chosen due to the interest of measuring the impact of home styling on the final price.

Final model 1

\[ Y_{\text{Final Price}_i} = \beta_0 + \beta_1 x_{\text{Start Price}_i} + \beta_2 x_{\text{Rooms}_i} + \beta_3 x_{\text{Floor}_i} + \beta_4 x_{\text{Fee}_i} + \beta_5 x_{\text{Location}_i} + \beta_7 x_{1900-1949} + \beta_8 x_{1950-1999} + \beta_9 x_{\text{homestyling}_i} + \epsilon_i, \]

Final model 2

\[ Y_{\text{Final Price}_i} = \beta_0 + \beta_1 x_{\text{Start Price}_i} + \beta_3 x_{\text{Floor}_i} + \beta_4 x_{\text{Fee}_i} + \beta_5 x_{\text{Location}_i} + \beta_7 x_{1900-1949} + \beta_8 x_{1950-1999} + \beta_9 x_{\text{homestyling}_i} + \epsilon_i, \]

Figure 7: Plot for Final model 1
Figure 8: Plots for Final model 2

<table>
<thead>
<tr>
<th>Models</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>AIC</th>
<th>BIC</th>
<th>Std.Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final 1</td>
<td>0.9354</td>
<td>0.9352</td>
<td>40849.41</td>
<td>40906.2</td>
<td>3072</td>
</tr>
<tr>
<td>Final 2</td>
<td>0.9353</td>
<td>0.9351</td>
<td>40849.53</td>
<td>40900.64</td>
<td>3073</td>
</tr>
</tbody>
</table>

Table 9: Statistics of the Final models

By looking at figure seven and eight it can be seen that they do not differ very much from the initial model and thus are still valid and normal.

Table nine shows that different statistic measures give different results regarding which of the models that is the most accurate. Thus, the results yield no unique model but instead two similar models.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Estimates</th>
<th>Std.Error</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5.426e+03</td>
<td>4.882e+02</td>
<td>4468.9886306</td>
<td>6383.6082500</td>
</tr>
<tr>
<td>Starting Price</td>
<td>9.800e-01</td>
<td>8.333e-03</td>
<td>0.9636176</td>
<td>0.9963007</td>
</tr>
<tr>
<td>Number of Rooms</td>
<td>-1.644e+02</td>
<td>1.131e+02</td>
<td>-386.1263284</td>
<td>57.3857089</td>
</tr>
<tr>
<td>Floor Level</td>
<td>3.897e+02</td>
<td>1.477e+02</td>
<td>100.0564668</td>
<td>679.3324902</td>
</tr>
<tr>
<td>Rental Fee</td>
<td>-4.608e-01</td>
<td>1.095e-01</td>
<td>-0.6756646</td>
<td>-0.2460238</td>
</tr>
<tr>
<td>Location</td>
<td>9.598e+02</td>
<td>2.690e+02</td>
<td>432.2777635</td>
<td>1487.239470</td>
</tr>
<tr>
<td>Construction 1900-1949</td>
<td>9.304e+02</td>
<td>1.648e+02</td>
<td>607.3064072</td>
<td>1253.4993661</td>
</tr>
<tr>
<td>Construction 2000-2017</td>
<td>-1.229e+03</td>
<td>1.817e+02</td>
<td>-1585.2976117</td>
<td>-872.5909571</td>
</tr>
<tr>
<td>home styling</td>
<td>7.110e+02</td>
<td>3.299e+02</td>
<td>64.0604845</td>
<td>1357.9803253</td>
</tr>
</tbody>
</table>

Table 10: Table of Confidence Intervals for Final model 1
<table>
<thead>
<tr>
<th>Variables</th>
<th>Estimates</th>
<th>Std.Error</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5.396e+03</td>
<td>4.878e+02</td>
<td>4439.3275464</td>
<td>6352.6961775</td>
</tr>
<tr>
<td>Starting Price</td>
<td>9.806e-01</td>
<td>8.325e-03</td>
<td>0.9642329</td>
<td>0.9968844</td>
</tr>
<tr>
<td>Floor Level</td>
<td>3.797e+02</td>
<td>1.476e+02</td>
<td>90.3098645</td>
<td>669.1083526</td>
</tr>
<tr>
<td>Rental Fee</td>
<td>-5.760e-01</td>
<td>7.568e-02</td>
<td>-0.7244051</td>
<td>-0.4275705</td>
</tr>
<tr>
<td>Location</td>
<td>9.602e+02</td>
<td>2.690e+02</td>
<td>432.5534291</td>
<td>1487.7813761</td>
</tr>
<tr>
<td>Construction 1900-1949</td>
<td>9.414e+02</td>
<td>1.646e+02</td>
<td>618.5968311</td>
<td>1264.2706010</td>
</tr>
<tr>
<td>Construction 2000-2017</td>
<td>-1.206e+03</td>
<td>1.811e+02</td>
<td>-1560.9341446</td>
<td>-850.7727674</td>
</tr>
<tr>
<td>home styling</td>
<td>7.228e+02</td>
<td>3.299e+02</td>
<td>75.8294820</td>
<td>1369.6953120</td>
</tr>
</tbody>
</table>

Table 11: Table of Confidence Intervals for Final model 2

The confidence interval in table ten and eleven for both of the Final models imply that home styling does have a positive effect on the final price. The width of the home styling intervals differ slightly between the models, though there is an observable difference of the magnitude. Final model 2 is of higher magnitude than model 1, which denotes that model 2 has a greater effect on the final price, considering home styling.
5 Discussion and Conclusion

A discussion of the results and the conclusions that the authors of the Bachelor of Science thesis have drawn are presented in this chapter. The conclusions are based from the analysis with the intention to answer the formulation of questions that are presented in the purpose and aim.

5.1 Discussion of the Results

The estimated coefficient ($\beta$) that represents home styling from final model 2 is $7.048e + 02$, which shows that home styling has a positive effect on the final price. The level of significance is 0.99 and the confidence interval is positive. This result shows that home styling can affect the final price of a condominium. When applying all possible regression, we can conclude that other variables have bigger impact on the final price than home styling, since home styling only is included in the model when allowing 7 or more variables. Starting price, construction year B (1900-1949) and location seems to affect the final price the most. Even though home styling is significant, other variables tend to affect more. The apartments already have high value to begin with so if the home styling is very expensive it may not be worth it, but this is an issue for further studies.

Since home styling is a dummy variable it is hard to tell how much it can affect the final price. As written in the introduction it is highly likely that the broker will put a higher starting price on the apartment to begin with if it has been styled. If so, the whole regression analysis is completely useless for the purpose of establishing a linear correlation between home styling and final prices. It is also hard to know if investing in home styling actually payoff. If the price of the home styling was available, a more thorough analysis could have been made. Given the price of different home stylings a break-even could be calculated.

This study can only determine if home styling does affect the final price or not, therefore it is hard to tell how big impact styling has on socioeconomics and society. As described in the introduction there are other factors that condominium prices depend on such as demand, supply, mortgage rates etc. and in the study it is also proven that other factors tend to affect the final price more. It is though, proven that home styling has some positive effect on a sale, and real estate agencies are investing more and more in home styling. In this moment we know that the housing
prices are rising, but it is hard to tell if home styling has an important role in this price rise. Home styling will additionally open up the possibility of new jobs and by that increase the employment and the economy.
6 Conclusion

The main purpose of this thesis was to determine whether or not home styling has a positive effect on the final price of condominiums in Uppsala. We also desired to answer four more targeted questions that are presented below.

- Is there a linear relationship between home styling and the response, final price?
- Is it scientifically recommended to invest in home styling based on our results?
- To what extent does home styling influence the final price?
- How can home styling contribute to the society and economy?

The linear relationship between the Final price and home styling is met due to the significance of the coefficient $\beta_{\text{homestyling}}$. This concludes that home styling has a notable effect on the final price. The results that were drawn from this thesis show that the final prices were positively affected by home styling. Although we did not manage to measure the influence of the effect and therefore a conclusion regarding whether it is scientifically advisable to invest in home styling can not be drawn.

As mentioned in the background and as proven in the results, home styling can increase the value of an object and in turn increase the profit for both the seller and the broker. The popularity of home styling is growing for each year as well as the housing prices in Sweden. If home styling keep on rising it may vitalize the rise in prices, though only with a modest extent.