



**MÄLARDALENS HÖGSKOLA  
ESKILSTUNA VÄSTERÅS**

Bachelor Thesis  
Spring 2020

# What impacts the formation of prices of apartments in Vasteras?

*With a hedonic pricing model approach*

Authors  
Hawta Lak  
Tamerlan Shikhalizade

Supervisor  
Clas Eriksson

# Abstract

Determining and predicting the exact prices of apartments is a complex task. It requires the data on the factors that directly influence the price of the apartment, such as the number of rooms or location of the apartment, and the information about the factors that indirectly affect the price, such as the availability of public transport and public goods near the apartment. One of the limitations of our work is the lack of data on the availability of indirect factors, and so in this paper we purely focus, and determine to what extent direct factors influence the formation of the final price. We find the influence of each of these factors with the help of the hedonic pricing, and the method of linear regression. After the first regression we identified which variable is least significant for our work and removed it. In our case it happened to be the variable Floor that identifies the level of the apartment. Further, we also test other types of regressions, such as semi – log regression, double – log regression, and quadratic regression. This is done to identify which of the regressions demonstrates the clearest picture on the effects of the variables. In other words, in which of the regressions the variables have the most significant parameter values. We found the Regression Five, a quadratic regression, to be an equation with the most significant parameter values. We also identified that the variables Rooms (indicating the number of rooms in the apartments) and Share (indicating the corporate share in the building) to have the biggest impact on the formation of the final price. Thus, we conclude that the variables Rooms and Share have the most significant influence on the price, whereas a quadratic regression (in this paper Regression Five) presents an equation with the most significant values of parameters and the highest degree of explanation.

<b>Introduction</b> .....	<b>5</b>
Background .....	5
Limitations of the Thesis .....	5
The problem .....	6
<b>Theory</b> .....	<b>7</b>
Consumer theory.....	7
Hedonic pricing .....	8
<b>Method</b> .....	<b>9</b>
Linear regression.....	9
Semi-log equation .....	9
Double log equation.....	9
Quadric equation .....	9
The Data .....	10
Summary Charts .....	11
Chart one.....	11
Chart two.....	11
<b>Regressions</b> .....	<b>12</b>
The First Regression .....	12
Expectations of the First Regression .....	12
Results of the First Regression .....	13
Interpretation of equation (1).....	14
The Second Regression .....	14
Expectations of the Second Regression .....	15
Results of the Second Regression .....	15
Interpretation of equation (2).....	16
The Third Regression.....	16
Expectations of the Third Regression.....	17
Results of the Third Regression.....	17
Interpretation of equation (3).....	18
The Fourth regression .....	18
Expectations of the Fourth Regression .....	19
Results of the Forth Regression .....	19
Interpretation of equation (4).....	20

Fifth regression.....	20
Expectations on the Fifth Regression .....	20
Results of the Fifth Regression .....	21
The interpretation of equation (5) .....	22
Regression six .....	22
Interpretation of regression six.....	23
Overview over all regressions .....	23
Chart Three .....	23
<b>Conclusion .....</b>	<b>24</b>
<b>References .....</b>	<b>26</b>

# Introduction

## Background

How is the price of an apartment formed? What affects the price of an apartment? And more precisely, how do different factors affect the price of an apartment. In this paper we are going to answer these questions and examine the key factors that affect the formation of apartment prices in the city of Vasteras, Sweden. This work is going to be based on the principle of hedonic pricing, and in order to see the impact of the influence and the effect of certain variables on the final price of the apartments, we are going to run a number of regressions. These regressions are going to tell us how each factor contributes to the formation of the final price. Moreover, we will try to identify which combination of variables best describes and explains the variation of the dependent variable. In addition, we will test how different variables affect each other. We will also discuss the importance of Limitations of the Thesis, as well as the concepts of Consumer Behavior and Public Goods. In this paper we will strongly rely on the empirical data, and several scientific works and papers.

## Limitations of the Thesis

This thesis is aimed on giving the closest and most accurate result on the impact of various factors on the formation of apartment prices in the city of Vasteras. However, there are a few limitations that make the final result not as perfect and accurate as we would want it to be.

The first big obstacle that we face is the lack of necessary data on houses. The original data set that we are working with has the statistics on the variables of 196 apartments and houses. However, the number of houses observed is only 33. Such relatively low number might be insufficient for an unbiased final result and research. This leaves us with no choice but to focus on apartments whose amount of observations are 163 which is sufficient.

In addition to that we lack information on some factors that affect the prices of apartments in other towns in Sweden. We believe that a comparison of the impact of factors in Vasteras, and the same factors in other cities, could be interesting and could show a certain common pattern in the influence of certain variables on the pricing of apartments amongst different cities, but this is not feasible.

Another limitation that we experience is the lack of information on the availability of elevators in the apartments. Later in the work we will see that this small limitation can affect the final prices of the apartments differently. The problem is that the variable “Floor” (which later will be introduced in more detail) is going to appear to be insignificant in the formation of the final price of the apartment. We believe that if we had sufficient information on which apartments have an elevator, we would be able to present a dummy variable, identifying which places are equipped with an elevator and hopefully this would add the variable Floor more significance. This is backed by the fact that if there is an elevator present, then living on the higher floors becomes a luxury, and so living on the higher floor positively affects the price. At the same, if there is no elevator in the building, living on the upper level will negatively affect the price. Unfortunately, as we lacked this information, we couldn’t adjust our regression to it.

Multicollinearity is another difficulty that we face. Certain variables have a high level of multicollinearity between each other, which strongly reduces the significance of some parameters. For this reason, we have to remove some variables. This was concluded after running a few regressions with certain variables and finding that they contributed to very insignificant results. These variables are not included in the final version of this paper.

The last limitation of this work is the Consumer Preferences. Later in this paper we will discuss the importance of Consumer Preferences and why is it vital to clearly understand this concept. We should understand that every person has their own personal choices and preferences. These preferences might potentially affect the formation of the final prices of the apartments. As we do not possess the information and data on personal preferences of all consumers, we cannot add them to our regressions or predict how exactly they might influence the final prices of apartments.

## The problem

*What variables effects the apartment prices the most in Vasteras?*

# Theory

## Consumer theory

Based on Kelvin J. Lancaster's work, Rosen (1974) developed an equation to describe the market equilibrium based on analysis of consumer and producer decisions. The following equation was developed:

$$Q_D = Q_S$$

In this equation the Market Quantity Demand ( $Q_D$ ) equals the Market Quantity Supply ( $Q_S$ ). Market equilibrium is a model where we can see how the price of a good or a service reacts to changes in the quantity supplied and demanded in the market. For the price to be at the equilibrium level, there are two conditions that must be satisfied. The first condition states that the quantity produced should equal the quantity consumed. The second condition states that the price should be at the level where the consumers are willing to pay for a good, and producers are ready to produce the good.

According to Lundmark (2010) consumer theory explains how a consumer spends his/her income on goods and services that gives the consumer most benefit.

The reason why the basic consumer theory is explained is because it is significant to understand the connection between the consumer theory, which is the traditional model and the new model which is called hedonic pricing that we are going to discuss later.

Axelsson, Holmlund, Jacobsson, Löfgren and Puu (1998) state that a limitation of the consumer choice is the level of income of a consumer. An important assumption to consumer theory is that a consumer always consumes goods and services that benefit them the most and never spends more than they can afford (their income). Another important assumption is that the consumer preferences are stable and are not random. This means the consumer does not behave randomly and it is possible to detect a behavioral pattern.

These traits of consumer behavior, described above, as well as described by Pindyck and Rubinfeld (2013) are called utility maximation. The graph for utility maximation is a downward sloping budget restriction (income), combined with multiple indifference curves. The indifference curve shows the combination of goods for which the consumer is indifferent. The consumer will always choose the indifference curve that tangents the budget line because it is the point of the optimal combination of goods. The traditional theory explains that the characteristics of a good do not cause the satisfaction of the consumer but only the good itself. Rosen (1974) does not agree with this and develops a new view on this.

## Hedonic pricing

Rosen (1974) explains that hedonic pricing is a method in economics that can be used to describe a pricing system or a model where the price of a product is not limited to the direct value of the object. According to this method there are internal and external factors that affect and shape the final price or the final value of an object. This method is often applied in the real estate market, but is not limited exclusively to it (Tyrväinen, 1997).

For example, suppose that a variable such as size of a house can be scaled from 1 to 5 and the ambition is to estimate the result of an increase in the size of a house in an area by one unit. From here it is possible to estimate the relationship between the price of a house and the size of the house, but there are two problems with it. The first problem is that the value of the individual house depends on other factors, such as living area which is most likely correlated with the size of the house. The living areas of apartments or houses in downtowns are probably smaller than the living areas of apartments or houses in the outskirts of the town. The second problem is that the consumers who live in big houses value the size more than the consumers who live in small houses. (Boardman, Greenberg, Vining and Weimer, 2018)

According to Boardman, Greenberg, Vining and Weimer (2018) the hedonic pricing method can solve these two problems with two steps. The first step estimates the relationship between an objects price or value and the objects variables that can influence the price or value. The next step estimates the consumers' willingness to pay for the variables of the object depending on their preferences. After these two steps with the help of hedonic pricing method, one can compute the change in consumer surplus depending on variables of an object. It is worth noting that in this paper we are going to focus predominantly on the first step of this method, due to the lack of data and available information necessary for the second method. A typical hedonic price function equation can be written as:

$$P = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

The dependent variable  $P$  is the price, while  $\beta_1$  and  $\beta_2$  are parameters. When  $X_1$  and  $X_2$  increase by one unit, the final price increases by  $\beta_1$  and  $\beta_2$  units, respectively.



# Method

## Linear regression

The method which we use to get the final results and investigate the problem is Linear Regression. According to Lewis-Beck (2015) linear regression is a common and useful method to analyze the relationship between variables. The regression consists of one dependent variable and one or more independent variables. When a regression analysis is performed it is important to look at the coefficient of determination, also called R-squared. Here R-squared measures how much of the total variation in the dependent variable can be explained by the independent variables. The measurements are between 0 and 1 and the closer to 1 the better. For example, if the R-squared is 0,97 it means that 97 percent of the total variation of the dependent variable around its mean is explained by the regression model. The R-squared are defined as:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

## Semi-log equation

According to Stock and Watson (2015) a semi-log function is an equation where some parts of the equation are in logarithmic form. There are no restrictions for which parts in the equation that can be in logarithmic form. Therefore, it can be either the dependent variable or one or some of the independent variables. The reason that the semi-log function form is used in a regression is to see how much one variable changes in percentage terms when other changes in straight units.

## Double log equation

Stock and Watson (2015) explain that the double log function is a common functional form. It is different from the linear regression model because in the linear model the slopes are constant, and the elasticities are not. It is the opposite in a double log function and that leads to the estimated coefficients can be interpreted as elasticities.

## Quadric equation

In the equation some of the variables are in a square form and that is why the quadratic equation model is appropriate to use. The model focuses on looking at when the independent variables change sign from negative to positive or the other way around. The variables change the sign resulting from an increase or decrease in the variables. (Stock and Watson, 2015)

## The Data

The data that we are using in this work is the list of apartment sales in Vasteras from 2016. We have received this data set from our supervisor. There are 163 observations in the dataset. The table below shows which variables have been chosen for the regressions. The left part in table are the shortenings of the variables and the right side provides a little more description of the variables. We picked these variables since we believed that they should influence the price the most and should yield to the most interesting results. In our work the variable “Heat” indicates if the heating is included in the fees of the apartment or not. It is a dummy variable and is equal to 1 if the heating is part of the fee. On the other hand, the variable “Fee” represents the extra monthly costs that the customer has to pay when owning the apartment (such as electricity, parking, heating and etc). “Rooms” shows the number of rooms in the apartment. “Year” reflects the year the apartment was built in (For Example: 1997, 2005 or 2012). Variable “Floor” represents the floor on which the apartment is located. It is also worth noting that the variable “Floor” is not continuous, as we only focus on the floor that the apartment is located on, and not the total height of the building. The variable “Share” stands for the housing cooperative share. Two variables that are related to the location of the apartment are “East” and” North” that are based on where the apartment is located. They represent the location of the apartment (whether the apartment is located relative closer to the east, or to the north in Vasteras). It is worth noting that the north part of the town is considered to be a less attractive location for households, as it is further from the lake, which is located on the south and which a lot of people would prefer to live next to.

<b>Heat</b>	Heating included. A dummy variable, equal to 1 if heat is included
<b>Fee</b>	Monthly fee
<b>Rooms</b>	Number of rooms
<b>Year</b>	The year the building was built
<b>Floor</b>	Apartment floor
<b>Share</b>	Housing cooperative share

<b>East</b>	Location rt90 geopt east (gps)
<b>North</b>	Location rt90 geopt north (gps)

## Summary Charts

Below are two charts of the data that is being used in the regressions. The reason for this is to give the reader a quick view and more info over the data. The first chart contains of minimum, maximum, average value and sigma. The second chart is a correlation matrix over variables that is being used in the regressions. By looking at the correlation matrix the most correlated variables are the dependent variable price and the variable share. This is not a surprise because the more area you own of a building the more it is going to cost you. The least correlation between variables are between year and floor. This outcome depends on there is no clear connection between which floor you live on and which year the house is built. There could be a stronger connection between these variables if the house was old and poorly built. It would maybe be seen as a danger to live on a higher floor if the house would come to fall apart. When buying a house with several floors it is taken for granted that every floor is livable no matter age of the house.

### Chart one

	<b>Min</b>	<b>Max</b>	<b>Average value</b>	<b>Sigma</b>
<b>Price</b>	300000	4500000	1233691.36	711040.58
<b>Heat</b>	0	1	0.75	0.43
<b>Fee</b>	1463	8715	3902.39	1304.81
<b>Rooms</b>	1	6	2.43	1.06
<b>Year</b>	1902	2014	1965.6	20.75
<b>Floor</b>	0	13	2.45	1.95
<b>Share</b>	0.0012	17.30	1.68	2.34
<b>East</b>	1529200	1546321	1542021.77	2460.72
<b>North</b>	6592731	6621487	6611295.72	2422.43

### Chart two

	Price	Heat	Fee	Rooms	Year	Floor	Share	East	North
Price	1	0.12	0.57	0.66	-0.16	0.28	0.79	-0.13	-0.17
Heat	0.12	1	0.11	0.07	0.02	0.07	0.15	-0.08	-0.18
Fee	0.57	0.11	1	0.75	0.15	0.18	0.56	-0.26	-0.22
Rooms	0.66	0.07	0.75	1	-0.05	0.19	0.42	-0.15	-0.13
Year	-0.16	0.02	0.15	-0.05	1	-0.34	-0.24	-0.14	-0.11
Floor	0.29	0.07	0.18	0.19	-0.34	1	0.34	-0.05	-0.06
Share	0.79	0.15	0.56	0.42	-0.24	0.34	1	-0.16	-0.14
East	-0.13	-0.08	-0.26	-0.15	-0.14	-0.05	-0.16	1	0.64
North	-0.17	-0.18	-0.22	-0.13	-0.11	-0.06	-0.14	0.64	1

## Regressions

### The First Regression

The first linear regression is used to see which of the independent variables have the most effect on the depended variable which is the price. The independent variables are heat, fee, rooms, year, floor, living area, housing cooperative share, east and north. The reason why we have so many different independent variables is to see how big of an impact each variable has on the price. The equation for the regression looks like:

$$Price = \beta_0 + \beta_1heat + \beta_2fee + \beta_3rooms + \beta_4year + \beta_5floor + \beta_6share + \beta_7east + \beta_8north + \varepsilon$$

### Expectations of the First Regression

The heat that is included in the rent is expected to have a positive effect on the price because people like to live in an apartment where they do not have scrimp on the heat. The monthly fee is most likely to be negative because people do not like to have high expenditures and therefore, it decreases the willingness of buying an apartment. When the monthly fee is increased it will thus have a negative impact on the price. Most consumers probably value the number of rooms very high and therefore is it expected to be positive.

The year indicates the year of building of the apartment. The higher is the year (the newer the apartment), the higher is going to be its price.

The coefficient of level of the floor of the apartment is probably going to be positive because you get a better view depending on the level. This of course depends on the availability of an elevator in the building. If there is an elevator present, then the higher floor is going to positively affect the price. On the other hand, if there is no elevator, then the higher floor is going to decrease the price. Unfortunately, we do not possess the data on the presence of elevator in any of the buildings, and so, it is very hard to draw any concrete assumptions and predictions upon this issue. Housing cooperative share are most likely to be positive because the bigger of share the higher the price. The independent variable east indicates the area of the town that has a relatively lower level of criminality, which results in a higher price of housing. The north is going to be negative as it indicated the further distance from the lake (in other words it is negative because more people prefer to live by the lake, and the further it is from the lake, the cheaper the price is expected to be).

## Results of the First Regression

After running the first regression presented above, by using the available data on the variables, these are the Estimates,  $t$  - Values and  $p$  - Values that we got. We also acquired the values for R-Squared and the Adjusted R-Squared.

	<b>Estimate</b>	<b>T-value</b>	<b>P-value</b>
<b>Intercept</b>	230755299.94	1.948	0.046*
<b>Heat</b>	31808.28	0.783	0.49
<b>Fee</b>	-157.59	-2.331	0.023*
<b>Rooms</b>	447586.939	5.810	3.09E-7*
<b>Year</b>	2927.63	1.083	0.28
<b>Floor</b>	2.884	0.273	0.78
<b>Share</b>	205627.630	8.720	5.12E-14*
<b>East</b>	22.520	0.873	0.3863
<b>North</b>	-29.142	-2.541	0.025*

<b>R-squared</b>	0.7845	<b>Adjusted R-squared</b>	0.7531
------------------	--------	---------------------------	--------

Variables with (\*) means that they are significant

The estimated equation (1):

$$\widehat{Price} = 230755299.94 + 31808.28heat - 157.59fee + 447586.939rooms + 2927.63year + 2.884floor + 205627.630share + 22.520east - 29.1426north$$

## Interpretation of equation (1)

From equation 1 we can conclude several things. Firstly, the biggest impact on the final price of the apartment is made by the housing cooperative share, as the bigger the share, the much higher the price of an apartment is. As predicted, the number of rooms is extremely valued among consumers too. Thus, every additional room in the apartment greatly increases the total price of the apartment. We can also see that heating is not significant and we cannot comment it. The level of the floor of the apartment also increases the price of the apartment, however not significantly. The year of the building also has a strong positive impact on the final price of the housing but is not significant. The variable east has a positive impact on the price, but this variable is also not significant. Whereas being located on the north decreases the price of the apartment.

We should also look at the R – Squared, which can be defined as the proportion of variance in the outcome variable which is explained by the predictor variables in the sample, according to Jeremy Miles (2005). As seen from the table above, R – Squared is equal to 0,7845, which is a relatively high result that indicated that most of the variation in the dependent variable can be explained by the independent variables.

## The Second Regression

The second regression is very similar to the first regression, but without the variable level of floor. The reason is that there is no variable for elevator in this dataset which it is believed to be an important variable when level of floor is included. The remaining seven variables were chosen because they were considered to be the most interesting after the first result. The second regression can be written as:

$$Price = \beta_0 + \beta_1heat + \beta_2fee + \beta_3rooms + \beta_4year + \beta_5share + \beta_6east + \beta_7north + \varepsilon$$

## Expectations of the Second Regression

The expectations of the second regression are similar to the expectations of the first regression. Thus, there are still beliefs that the housing cooperative share and number of rooms have positive impact on the price for the similar reasons as explained previously. If the factors housing cooperative share and number of rooms increase with one unit, the price will also increase. The monthly fee is also expected to have the same effect as before, and so we are expecting a negative influence on the price.

## Results of the Second Regression

	<b>Estimate</b>	<b>T-value</b>	<b>P-value</b>
<b>Intercept</b>	150980133.84	1.978	0.045*
<b>Heat</b>	30259.47	0.257	0.79*
<b>Fee</b>	-143.87	-2.136	0.022*
<b>Rooms</b>	401069.77	5.814	2.63E-7*
<b>Year</b>	2747.63	1.062	0.29
<b>Share</b>	207650.61	8.971	1.28E-14*
<b>East</b>	32.79	1.554	0.13
<b>North</b>	-43.34	-2.245	0.028*
<b>R-squared</b>	0.7776	<b>Adjusted R-squared</b>	0.7513

*Variables with (\*) means that they are significant*

*The estimated equation (2):*

$$\widehat{Price} = 150980133.84 + 30259.47heat - 143.87fee + 401069.77rooms + 2747.63year + 207650.61share + 32.79east - 43.346north$$

## Interpretation of equation (2)

There are some changes in the estimated second regression compared to the first one. The value of the R – Squared in regression two has dropped in comparison to the value of R – Squared in regression one (0.7776 versus 0.7845 respectively). This implies that slightly less variations in the estimated dependent variable can be explained by the estimated independent variables. If we only base our conclusions on the value of R-Squared then it is safe to say that the second regression is hardly different from the first one, experiencing a slight decrease in its value. It is important to also compare the adjusted R – Squared values of the two regressions. Adjusted R – Squared can be defined as a modified version of a normal R – Squared, that has been adjusted to the number of variables in the model. As seen from the tables, there is hardly any difference in the values of the adjusted R – Squared (0.7531 in the first regression to 0.7513 in the second regression). Based on this we can say that these seven variables are the most interesting to work with (level of floor excluded), and so furthermore we will only proceed working with these seven variables.

The variable heat has a positive impact on the price but is not significant and cannot be commented on. Drawing the parallels with the results of Regression one, in Regression two, the impact of Share has increased, as the Estimated Value of this parameter has risen from 205627.630 to 207650.61. The variable Fee, however, has slightly decreased (in absolute value) from -157.59 to -143.87. The change is not much which means that it was a good thing to drop the floor variable. The variable Rooms has the largest impact on the dependent variable. The reason behind it might be attached to a fact that a higher living area, that usually results from the higher number of rooms, leads to higher prices for apartments. If the variable increases by one unit, the price will also increase by 401069.77 SEK. The variable year also has a positive impact on the dependent variable but is still insignificant. The two variables north and east have different impacts on the price. East is not significant while north is. If the variable north increases with one unit, the price will decrease with 43.34 SEK

## The Third Regression

The third regression is a little bit different from the previous regressions since now the dependent variable is in a logarithmic form while the independent variables are not. This kind of regression is called a semi-log regression model. The equation looks like this:

$$\text{LogPrice} = \beta_0 + \beta_1 \text{Heat} + \beta_2 \text{Fee} + \beta_3 \text{Rooms} + \beta_4 \text{Year} + \beta_5 \text{Share} + \beta_6 \text{East} + \beta_7 \text{North} + \varepsilon$$



## Expectations of the Third Regression

The expectations of the independent variables will not differ from the expectations on the ones in the second regression since there are no changes in them, and they are not in logarithmic forms either. The changes in the final price, resulting from increases of the independent variables will be measured in percent. That is because the dependent variable is expressed in a logarithmic form.

## Results of the Third Regression

	<b>Estimate</b>	<b>T-value</b>	<b>P-value</b>
<b>Intercept</b>	178.66666858	2.109	0.039*
<b>Year</b>	0.00381978	1.976	0.052
<b>Heat</b>	0.05467312	0.618	0.54
<b>Share</b>	0.10380971	6.271	4.58E-7*
<b>East</b>	0.00001502	0.996	0.32
<b>North</b>	-0.00002963	-2.147	0.035*
<b>Fee</b>	-0.00011548	-2.388	0.02*
<b>Rooms</b>	0.32489202	6.585	1.36E-7*
<b>R-squared</b>	0.7122	<b>Adjusted R-squared</b>	0.6781

*Variables with (\*) means that they are significant*

*The estimated equation (3):*

$$\widehat{\text{LogPrice}} = 178.66666858 + 0.00381978\text{year} + 0.05467312\text{heat} + 0.10380971\text{share} + 0.00001502\text{east} - 0.00002963\text{north} - 0.00011548\text{fee} + 0.32489202\text{rooms}$$

### Interpretation of equation (3)

Looking at R – Squared we can see that it is 0.7122 which is slightly lower than the value of R – Squared in the second regression, however we cannot compare them since price in the Third Regression is expressed in the logarithmic form. The R-Squared in Regression Three is quite strong and explains a lot of the variations in the logarithm of the dependent variable. The adjusted R – Squared here is also slightly lower than the adjusted R – Squared in the First Regression. We can conclude that the semi-log functional form is an appropriate model when it comes to analyzing the formation of apartment prices.

From the results we can see that the variable year still has a little positive impact on the price but still is not significant. Just like in the first and second regressions, the variable Fee has a negative sign and the price for the apartment drops by 0.00011548 percent with an increase of every unit in this explanatory variable. Just like in the earlier regressions, the variable Number of Rooms has a huge impact on the formation of the final price, with the price increasing by 0.32489202 with every room added. The value of the Share is also significant in this regression, just as in the previous regressions. It results in the increase of the final price of the apartment by 0.10390971 percent with an increase of one unit in Share. Similarly, to all other parameters that stay loyal to their signs, East and North also keep their original signs. North is still the one who is significant, resulting a decrease in the price 0.00002963 percent, with an addition of an extra unit. The variable heat, as expected it has a positive impact on the dependent variable but still is not significant.

### The Fourth regression

The fourth regression differs the most from other regressions. Here, both independent and dependent variables are expressed in a logarithmic form. The variables Year and Heat are not expressed in the logarithmic forms since you cannot have dummy variables expressed in a logarithmic form. The variables east and north are also not in logarithmic forms due to them performing better when they are in a normal (not logarithmic) state. The fourth regression can be written as:

$$\text{LogPrice} = \beta_0 + \beta_1\text{Year} + \beta_2\text{Heat} + \beta_3\text{LogShare} + \beta_4\text{East} + \beta_5\text{North} + \beta_6\text{LogFee} + \beta_7\text{LogRooms} + \varepsilon$$

## Expectations of the Fourth Regression

The expectations of Regression Four remain the same due to all factors remaining the same with the only difference of those variables being in a logarithmic form. When the equation is measured with percentages on both sides, the coefficients are interpreted as elasticities.

## Results of the Forth Regression

	<b>Estimate</b>	<b>T-value</b>	<b>P-value</b>
<b>Intercept</b>	216.97863007	2.032	0.041*
<b>Year</b>	0.00010922	0.617	0.54
<b>Heat</b>	0.03252734	0.309	0.76
<b>Log Share</b>	0.05685393	2.901	0.005*
<b>East</b>	0.00002862	0.894	0.37
<b>North</b>	-0.00003729	-1.810	0.047*
<b>Log Fee</b>	-0.13240977	-1.995	0.047*
<b>Log Rooms</b>	0.78445544	4.637	2.01E-5*
<b>R-squared</b>	0.5936	<b>Adjusted R-squared</b>	0.5453

*Variables with (\*) means that they are significant*

*The estimated equation (4):*

$$\widehat{\text{LogPrice}} = 216.97863007 + 0.00010922\text{Year} + 0.03252734\text{Heat} \\ + 0.05685393\text{LogShare} + 0.00002862\text{East} - 0.00003729\text{North} \\ - 0.13240977\text{LogFee} + 0.78445544\text{LogRooms}$$

## Interpretation of equation (4)

The R – Squared is lower in this regression than in the second regression. However, we cannot compare the two results due to the differences in the forms of the equations.

The variable Year is not significant and will therefore not be commented. Heating still has a positive impact on the price but is also insignificant. Log Share has a positive impact on the price, as when Share increases by one percent the price is going to increase by 0.05685393 percent. Log Rooms is still strong in this regression, similarly to how it acts in all other regressions we investigated before. An increase by one percent of this variable will result in an increase by 0.78844544 percent of the final price. East and North act similarly to how they performed before where east is not significant, and north is. Just as was predicted originally, and achieved in most regressions, the variable log Fee has a negative sign. An increase by one percent of this variable results in the drop of the price by -0.13240977 percent.

## Fifth regression

In this regression the dependent variable is still price. There are three additional independent terms that are in a squared form: Rooms, Fee and Share. Just as the previous regression some of the independent variables are not in a squared form for the same reasons as before. The variables that are in squared form are chosen because they are believed to be the most interesting. The fifth equation can be written as:

$$Price = \beta_0 + \beta_1heat + \beta_2year + \beta_3rooms + \beta_4rooms^2 + \beta_5fee + \beta_6fee^2 + \beta_7share + \beta_8share^2 + \beta_9east + \beta_{10}north + \varepsilon$$

## Expectations on the Fifth Regression

In this regression it is interesting to see the relationships between the squared variable and the unsquared variable to see if there are any turning points. For instance, this could be when a variable that has a negative impact becomes positive. The expectations of the independent variables that are not in a squared form are still the same as before. Monthly fee in the squared form is expected to be positive, so that the total effect becomes positive at some point. The turning point may not happen, but it is still useful to use because it brings in some non-linearities. The reason for this is that after some time the price will be zero if the monthly fee continues to grow, if other independent variables are held constant.

However, we know that the price cannot be zero. The variable share in the squared form is expected to have a positive impact on the price because the more area of the building the consumer owns the higher is the price. Rooms in the squared form is also expected to impact the price in a positive way. This might be due to so called "status effects". However, it is expected to not have the same amount of impact on the price as the unsquared form because it is believed that after some point the number of rooms will satisfy the consumer need.

## Results of the Fifth Regression

	<b>Estimate</b>	<b>T-value</b>	<b>P-value</b>
<b>Intercept</b>	239084586.73977	2.147	0.036*
<b>Year</b>	3288.60944	1.258	0.21
<b>Heat</b>	11754.71939	0.101	0.92
<b>Share</b>	1290243.67768	2.652	0.01*
<b>Share<sup>2</sup></b>	5125.75842	2.594	0.01*
<b>East</b>	22.64606	1.123	0.27
<b>North</b>	-42.20709	-2.327	0.023*
<b>Fee</b>	-798.10026	-3.328	0.0016*
<b>Fee<sup>2</sup></b>	0.06405	2.733	0.0083*
<b>Rooms</b>	786644.52105	3.398	0.0012*
<b>Rooms<sup>2</sup></b>	45906.71443	2.427	0.016*
<b>R-Squared</b>	0.8138	<b>Adjusted R-Squared</b>	0.7806

*Variables with (\*) means that they are significant*

The estimated equation (5):

$$\widehat{Price} = 239084586.73977 + 11754.71939heat + 3288.60944year + 786644.52105rooms + 45906.71443rooms^2 - 798.10026fee + 0.06405fee^2 + 1290243.67768share + 5125.75842share^2 + 22.64606east - 42.20709north$$

## The interpretation of equation (5)

The R-Squared in this regression equals to 0.8138, which is the highest R – Squared experienced so far in this work. This shows that such non-linear form of the equation explains most of the variation of the dependent variable. The fifth regression could be argued to be the best fit for the dataset if you base your opinion purely on the R-Squared. However due to the variations in some of the regressions it is not possible to determine this. The reason for this is that the third and fourth regression have different prerequisites because of the different functional forms in them. It is possible to compare the first and second regression to the fifth because they have the same form of the dependent variable.

The variable Heat still is not significant. The variable Year slightly increases the price by 3288.60944 units with every year but recall that this effect is not statistically significant. As always, Rooms has a huge impact on the price. This time the price increases by 832551.2355 (786644.52105 + 45906.71442)units when the variable Rooms increase with one unit (if the derivative is evaluated at the point where this variable is equal to one). If the apartment is large, an additional room increases the price more than it would do if the house were small. The parameter Fee has a negative impact on the dependent variable. If Fee increases by one unit the price will decrease by 798.03621 (0.06405 – 798.10026) units. As expected, Fee does not have a positive impact on the final price up until a certain point (which may be outside the sample). Share also has a positive impact on the price. The price increases by 1295369.435 (1290243.677 + 5125.75842) units when share increase by one unit. Just like Rooms, Share also experiences an increase in impact on the price after a certain point. East is still not significant. North has negative impact on the price.

## Regression six

In the previous regressions the intercept has been extremely high and therefore not quite believable. In this regression the aim is to make the intercept more realistic. The dependent variable is price and the independent variable is the number of rooms. By removing the independent variables east and north we can see the impact they had on the intercept.

	Estimate	T-value	P-value
Intercept	381389	3.155	0.0019*

<b>Rooms</b>	350776	7.707	1.35E-12*
<b>R-Squared</b>	0.2717	Adjusted R-Squared	0.2671

## Interpretation of regression six

As we can see the intercept is much lower than before, it has decreased to 383389 from 230755299. It means that the price for an apartment is 383389 SEK when all the independent variables are zero. The reason for this is that the variables east and north are measured in numbers and not in longitude and latitude. Because of the numbers are high in the variables east and north it makes the intercept much higher when they are included. For example, an observation can be 6611486 and 6611499 in the variable north. It is almost the same, but it differs 13 units between the apartments. Because of the high number in the variable north the intercept can be seen as unrealistic for the reason that it can increase with a couple of millions in price when in reality it is about hundred meters between them. It is not normal for an apartment with zero independent variables to cost hundreds of millions.

## Overview over all regressions

This is an overview of all five regressions that is been used in the result part. The number in are the estimates of the variables in the regressions. The variables that are significant is marked with a (\*). In this model we can get a clear view of all regressions.

### Chart Three

	<b>Regression one</b>	<b>Regression two</b>	<b>Regression three</b>	<b>Regression four</b>	<b>Regression five</b>	<b>Regression six</b>
<b>Dependent variable</b>	Price	Price	Log Price	Log Price	Price	Price
<b>Intercept</b>	230755299.94*	150980133.84*	178.666668558*	216.97863007*	239084586.73977*	381389*
<b>Heat</b>	31808.28	30259.47	0.05467312	0.03252734	11754.71939*	
<b>Fee</b>	-157.59*	-143.87*	-0.00001502*		798.10026*	

<b>Log Fee</b>				- 0.1324097 7*		
<b>Fee<sup>2</sup></b>					0.06405*	
<b>Rooms</b>	447586.93 9*	401069.77 *	0.32489202 *		786644.5210 5*	350776*
<b>Log Rooms</b>				0.7844554 4*		
<b>Rooms<sup>2</sup></b>					45906.71443 *	
<b>Year</b>	2927.63	2747.63	0.00381978	0.0001092 2	3288.60944	
<b>Floor</b>	2.884					
<b>Share</b>	205627.63 0*	207650.61 *	0.10380971 *		1290243.677 68*	
<b>Log Share</b>				0.0568539 3*		
<b>Share<sup>2</sup></b>					5125.75842*	
<b>East</b>	22.520	32.79	0.00001502	0.0000286 2	22.64606	
<b>North</b>	-29.14*	-43.34*	- 0.00002963 *	- 0.0000372 9*	-42.20709*	
<b>R-Squared</b>	0.7845	0.7776	0.7122	0.5936	0.8138	0.2717
<b>Adjusted R-Squared</b>	0.7531	0.7513	0.6781	0.5453	0.7806	0.2671

## Conclusion

Calculating and predicting the prices of apartments is a complicated challenge. Indeed, it is relatively easy to predict an approximate price, based on the location of the place, the number of rooms, floors and other factors, but to give a precise and accurate answer is much harder. The reason for that is an enormous number of unknown external factors, together with the sometimes-unpredictable behavior of the inner factors, some of which have just been mentioned above.



Nevertheless, we decided to take on this tough challenge, and explore the issue with the hedonic approach. We ran several regressions, the first of which was a basic regression displaying the effects of all the variables we were testing. It was important to identify the variables with the biggest influence and were the most significant, which happened to be Corporate Share and Number of Rooms.

One of the biggest problems that we have experienced throughout this work was multicollinearity. That's why, in order to see how certain variables would act in the absence of other variables, we ran two more regressions in order to see if we can get a different perspective on the influence of the factors. Indeed, certain parameters have slightly deteriorated or increased in their values but remained with the same sign and had a very similar proportional influence on the final price, to what they had before. This is good, as it demonstrates low levels of potential multicollinearity. One of the biggest problems with multicollinearity was avoided by leaving out the variable "Living area", which strongly correlated with the variable "Rooms". Furthermore, we ran a few more regressions to see if we get any differences in results. These regressions included, testing for Semi Log, Log and also squaring certain factors to see if they yield a different result with a positive sign, respectively. All of these tests have helped us to see better what, and to which extent, influences the creation of final prices for apartments in the city of Vasteras. We can conclude the non-linear Regression Five is the most preferred way of determining the final price of the apartments, as it yields the parameters with the highest level of significance, as well as produces an Adjusted R – Squared that explains the most variations in the estimated dependent variable amongst all tested regressions.

# References

Allison, Paul D. 1999. *Multiple regression*. SAGE Publications Inc.

Axelsson, Roger; Holmlund, Roger; Jacobsson, Roger; Löfgren, Karl-Gustaf; Puu, Tönu. 1998. *Mikroekonomi*. Second edition. Studentlitteratur AB, Lund.

Boardman, Anthony E; Greenberg, David H; Vining, Aiden R; Weimer, David L. 2018. *Cost-benefit analysis*. Fifth edition. Cambridge University Press.

Grunberg, Isabelle; Marc Stern, Inge Kaul. 1999. *Global Public Goods: International Cooperation in the 21st Century*. First edition. Oxford University Press.

Lewis-Beck, Colin. 2015. *Applied regression*. SAGE Publications Inc.

Lundmark, Robert. 2010. *Mikroekonomi, teori och tillämpning*. Studentlitteratur AB, Lund.

Miles, Jeremy. 2005. *Wiley StatsRef: Statistics Reference Online*. © John Wiley & Sons, Ltd. Abstract.

Pindyck, Robert S; Rubinfeld, Daniel L. (2013). *Microeconomics*. Eighth edition. Boston: Pearson Education limited.

Rosen, Sherwin. 1974. *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition*. *The Journal of Political Economy*, Volume 82, Issue 1, 34-55.

Sheather, Simon J. 2010. *A modern approach to regression with R*. Springer-Verlag New York Inc.

Stock, James H; Watson, Mark W. 2015. *Introduction to econometrics*. Third edition. Pearson Education Limited.

Tyrväinen, Liisa. 1997. *The amenity value of the urban forest: an application of the hedonic pricing method*. *Landscape and Urban Planning*, Volume 37, (Issues 3-4): p. 211–222.