DEGREE PROJECT, IN APPLIED MATHEMATICS AND INDUSTRIAL ECONOMICS , FIRST LEVEL
STOCKHOLM, *SWEDEN 2014*

Modelling Apartment Prices with the Multiple Linear Regression Model

ALEXANDER GUSTAFSSON, SEBASTIAN WOGENIUS

KTH ROYAL INSTITUTE OF TECHNOLOGY

SCI SCHOOL OF ENGINEERING SCIENCES

Modelling Apartment Prices with the Multiple Linear Regression Model

A L E X A N D E R G U S T A F S S O N
S E B A S T I A N W O G E N I U S

Degree Project in Applied Mathematics and Industrial Economics (15 credits)
Degree Progr. in Industrial Engineering and Management (300 credits)
Royal Institute of Technology year 2014
Supervisor at KTH was Tatjana Pavlenko
Examiner was Tatjana Pavlenko

TRITA-MAT-K 2014:06
ISRN-KTH/MAT/K--14/06--SE

Royal Institute of Technology
School of Engineering Sciences

KTH SCI
SE-100 44 Stockholm, Sweden

URL: www.kth.se/sci

Modelling Apartment Prices with the Multiple Linear Regression Model

Abstract

This thesis examines factors that are of most statistical significance for the sales prices of apartments in the Stockholm City Centre. Factors examined are *address, area, balcony, construction year, elevator, fireplace, floor number, maisonette, monthly fee, penthouse and number of rooms*. On the basis of this examination, a model for predicting prices of apartments is constructed. In order to evaluate how the factors influence the price, this thesis analyses sales statistics and the mathematical method used is the multiple linear regression model. In a minor case-study and literature review, included in this thesis, the relationship between proximity to public transport and the prices of apartments in Stockholm are examined.

The result of this thesis states that it is possible to construct a model, from the factors analysed, which can predict the prices of apartments in Stockholm City Centre with an explanation degree of 91% and a two million SEK confidence interval of 95%. Furthermore, a conclusion can be drawn that the model predicts lower priced apartments more accurately. In the case-study and literature review, the result indicates support for the hypothesis that proximity to public transport is positive for the price of an apartment. However, such a variable should be regarded with caution due to the purpose of the modelling, which differs between an individual application and a social economic application.

Modellering av lägenhetspriser med multipel linjär regression

Sammanfattning

Denna uppsats undersöker faktorer som är av störst statistisk signifikans för priset vid försäljning av lägenheter i Stockholms innerstad. Faktorer som undersöks är *adress, yta, balkong, byggår, hiss, kakelugn, våningsnummer, etage, månadsavgift, vindsvåning och antal rum*. Utifrån denna undersökning konstrueras en modell för att predicera priset på lägenheter. För att avgöra vilka faktorer som påverkar priset på lägenheter analyseras försäljningsstatistik. Den matematiska metoden som används är multipel linjär regressionsanalys. I en mindre litteratur- och fallstudie, inkluderad i denna uppsats, undersöks sambandet mellan närhet till kollektivtrafik och priset på lägenheter i Stockholm.

Resultatet av denna uppsats visar att det är möjligt att konstruera en modell, utifrån de faktorer som undersöks, som kan predicera priset på lägenheter i Stockholms innerstad med en förklaringsgrad på 91 % och ett två miljoner SEK konfidensintervall på 95 %. Vidare dras en slutsats att modellen preciderar lägenheter med ett lägre pris noggrannare. I litteratur- och fallstudien indikerar resultatet stöd för hypotesen att närhet till kollektivtrafik är positivt för priset på en lägenhet. Detta skall dock betraktas med försiktighet med anledning av syftet med modelleringen vilket skiljer sig mellan en individuell tillämpning och en samhällsekonomisk tillämpning.

Contents

List of Tables

List of Figures

1 Introduction

The demand for apartments in Stockholm is high and there is a lack of housing, especially apartments. In December 2013, Statistics Sweden (Swedish: Statistiska Centralbyrån) published an article *Stockholm citizens thrive despite lack of housing* (SCB 2013) on this issue referring to a report *Stockholm Country's housing market 2013* (Blume, Streiler, and Weston 2013) stating that there is a need for 6,000 more apartments per year in Stockholm. This is a significant amount compared to the the current construction rate of 10,000 apartments per year. There is an ongoing debate about how many apartments that should be built in Stockholm and where they should be located. (Jennische 2014) The yearly housing survey done by Länsstyrelsen Stockholm states that all municipals in Stockholm estimates a lack of housing. (Enheten för samhällsplanering 2014)

People value different things and according to the article *Stockholm citizens thrive despite lack of housing* (SCB 2013) people in Stockholm value proximity to public transportation to a great extent. A minor case-study and literature review, included in this thesis, focuses on examining if this aspect is reflected by the sales prices of apartments.

During the period December 2013 to February 2014, 1,521 apartments were sold in the Stockholm City Centre for a total value of 6.3 billion SEK. (Svensk Mäklarstatistik AB 2014) These amounts along with the above questions makes it interesting from a social economic perspective to know what makes an apartment valuable.

Similar studies have been done before using different data and statistics. The thesis *Estimation of apartment prices in the inner city of Stockholm using multiple regression analysis* (R. Gunnvald and P. Gunnvald 2012) suggests for further studies to also include a variable incorporating *"proximity of public transport"*, pointing out that such a variable would reflect how well an apartment is located. Furthermore, the paper *The Impact of Bus Rapid Transit and Metro Rail on Property Values in Guangzhou, China* (Salon 2014) states that

> "...proximity to the Metro and the Bus Rapid Transit have a substantial and statistically significant effect on apartment prices that varies by district and amenities provided..."

The article *The relationship between property values and railroad proximity: a study based*

on hedonic prices and real estate brokers' appraisals (Strand and Vågnes 2001) suggests a similar pattern for Oslo, Norway.

In this thesis a mathematical approach is applied to analyse sales statistics using the multiple regression model in order to construct a model that predicts the value of an apartment. The sales statistics are based on apartments sold in the Stockholm City Centre for the years 2012 and 2013, and incorporates the following variables: *address, area, balcony, construction year, elevator, fireplace, floor number, maisonette, monthly fee, penthouse and number of rooms.*

The problem statement of this thesis can be divided in two questions:

> *1. What factors are important when valuing an apartment and to what extent is it possible to predict the value of an apartment?*
>
> *2. Is proximity to public transportation an important aspect when valuing an apartment?*

The thesis is divided in three parts:

1. First part of the thesis focuses on explaining the mathematical theory behind the multiple regression model. This part is represented by chapter 2.

2. In the next part, an application of multiple regression is done on sales prices of apartments. This part is represented by chapters 3, 4 and 5.

3. At last a case-study and literature review is conducted in order to examine if proximity to subway stations can be used to improve the model for the value of an apartment. This part is represented by chapter 6.

Limits in this thesis are due to the data. For the second part, the application of the sales prices is limited to the variables that are available in the data. And for the third part, data over proximity to subway stations limits the study to a specific district in the Stockholm City Centre.

2 Background: The Multiple Regression Model

The basic model for econometric work is the linear regression model. It is an approach for modelling the relationship between a dependent variable and one or more explanatory variables, which will be referred to as covariates in this thesis. (Lang 2013, p. 18)

Linear regression can be used to fit a predictive model to a set of data values as well as a structural interpretation which allows for hypotheses testing. Structural interpretation means that we consider the covariates to influence the dependent variable, but not the other way round. (Lang 2013, p. 19)

This thesis will use the multiple regression model, which is valid when five basic assumptions are met. When these assumptions are met the ordinary least squares (OLS) estimator is guaranteed to be the optimal estimator. (Kennedy 2008, p. 40)

2.1 Definition and Terminology

The specification for the linear regression model is

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + e_i \qquad i = 1, 2, \ldots, n \tag{2.1}$$

In the expression, y_i is regarded as the dependent variable whose value depends on the covariates $x_{\bullet j}$. The parameters β_j are unknown, as is typically the variance, and are to be estimated from data. The error terms are normally distributed and denoted as e_i. (Lang 2013, pp. 18-19)

It is often more convenient to employ matrix notation:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e} \tag{2.2}$$

$\mathbf{Y}$ is a $n \times 1$ vector:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$\mathbf{X}$ is a $n \times (k+1)$ matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

β is $(k+1) \times 1$ vector:

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

$\mathbf{e}$ is $n \times 1$ vector:

$$\mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

2.1.1 Dummy Variables and Benchmarks

A dummy variable is an artificial variable constructed in order to take the value one whenever the phenomenon it represents occurs, and zero otherwise. It is used in the multiple linear regression model just like any other covariate. (Kennedy 2008, p. 232)

Benchmarks are used to make it easier to compare different dummy variables to the benchmark and to get round multicollinearity. (Lang 2013, p. 19) See section 2.5 for more information about multicollinearity.

2.2 Important Assumptions

The multiple linear regression model consists of five basic assumptions concerning the way in which the data are generated. (Kennedy 2008, p. 41)

1. The first assumption is that the dependent variable can be calculated as a linear function of the covariates, plus an error term. Thus, it should have the form of equation 2.1 or expressed with matrix notation as equation 2.2. (Kennedy 2008, p. 41)

   Violations of this assumption:

   - **Wrong regressors** - absence of relevant covariates and presence of irrelevant covariates.
   - **Nonlinearity** - the relationship between the dependent variable and the covariates is not linear.

2. The second assumption is that the expected value of the error term is zero, which can be expressed mathematically as $E[\mathbf{e}] = \mathbf{0}$. An estimator with the expected value of zero is called unbiased. (Kennedy 2008, p. 41)

3. The third assumption is that the error terms all have the same variance and are not correlated with one another. (Kennedy 2008, p. 41)

   Violations of this assumption:

   - **Heteroscedasticity** - the error terms do not have the same variance. Further explained in section 2.4.

4. The fourth assumption is that the covariates can be considered fixed in repeated samples, which means it is possible to redraw the sample with the same values for the covariates. This can be expressed mathematically as $E[\mathbf{e}\mathbf{e}^\top] = \sigma\mathbf{I}$. (Kennedy 2008, p. 41)

   Violations of this assumption:

   - **Errors in variables** - errors in measuring the covariates.
   - **Autoregression** - using a lagged value of the dependent variable as a covariate.

5. The fifth assumption is that the number of dependent variables is greater than the number of covaraites and that there are no exact linear releationship between the covariates. This implies that $rank\mathbf{X} \leq n$. (Kennedy 2008, p. 42)

   Violation of this assumption:

   - **Multicollinearity** - two or more covariates are approximately linearly correlated in the sample data. Further explained in section 2.5.

2.3 Ordinary Least Squares Estimation

The ordinary least square (OLS) estimator is considered the optimal estimator of the unknown parameters β when the assumptions of the multiple linear regression model are met. (Kennedy 2008, p. 40) The estimates of the OLS is denoted with a hat; e.g., the OLS of β is expressed as $\hat{\beta}$. (Lang 2013, p. 21)

The estimated $\hat{\beta}$ achieved by this method minimizes the sum of the squared errors. This is done by putting the derivative of the sum of the squared errors with respect to $\hat{\beta}$ equal to zero. (Lang 2013, p. 21)

The sum of the squared errors:

$$\begin{aligned}\sum_{i=1}^{n} \hat{e_i}^2 &= \sum_{i=1}^{n} (y_i - \hat{y_i}) \\ &= (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{X}\hat{\beta} - \hat{\beta}^\top \mathbf{X}^\top \mathbf{y} + \hat{\beta}^\top \mathbf{X}^\top \mathbf{X}\hat{\beta}\end{aligned}$$

The derivative with respect to β:

$$\begin{aligned}\frac{\partial(\mathbf{y}^\top \mathbf{y} - \mathbf{X}\hat{\beta} - \hat{\beta}^\top \mathbf{X}^\top \mathbf{y} + \hat{\beta}^\top \mathbf{X}^\top \mathbf{X}\hat{\beta})}{\partial \hat{\beta}} &= \mathbf{0} \\ -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\hat{\beta} &= \mathbf{0} \\ \mathbf{X}^\top \mathbf{y} &= \mathbf{X}^\top \mathbf{X}\hat{\beta}\mathbf{y}\end{aligned}$$

Hence, it follows that

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{2.3}$$

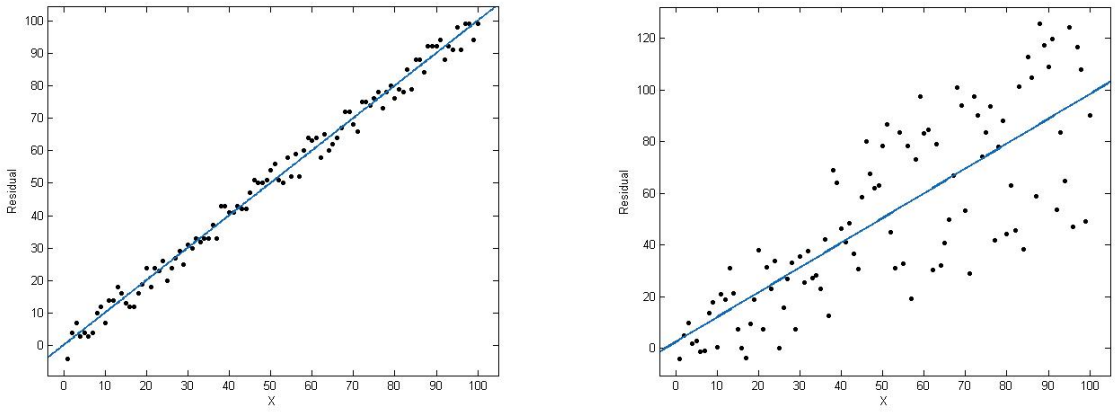
Under the multiple linear regression model's assumptions the OLS method is unbiased and thus $E(\hat{\beta}) = \beta$. The covariance of the OLS is calculated as (Lang 2013, p. 21)

$$Cov(\hat{\beta} \mid \mathbf{X}) = (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2 \tag{2.4}$$

2.4 Homoscedasticity and Heteroscedasticity

The third assumption states that the error terms all have the same variance. This is called homoscedasticity and may in mathematical terms be written as $Var(e_i|x_i) = \sigma^2$, where e_i is the error term and x_i is the measure of some covariate. An example of homoscedasticity is shown in figure 2.1.

The opposite of homoscedasticity is the phenomenon of heteroscedasticity, where the error term can be formulated as a function of x_i; for example, the error term increases for larger measurements of x_i. This can be described in mathematical terms as $Var(e_i|x_i) = f(x_i)$ and is shown in figure 2.1.



(a) Homoscedasticity

(b) Heteroscedasticity

Figure 2.1: Examples of residuals that are homoscedastic and heteroscedastic relative to some covariate x.

Heteroscedasticity is undesirable since it implies that the model is not as accurate for all input data. This inaccuracy occurs because the variance in the error term is not constant. As the example in figure 2.1 indicates: the error terms are greater for larger measurements of the covariate x. In order to confirm that assumption three is valid, it needs to be verified that heteroscedasticity is not a present issue for each covariate that is not a dummy variable.

2.4.1 Detecting Heteroscedasticity

There are various ways of detecting heteroscedasticity. We will present two, which are relevant for this thesis, in chronological order of there use.

Eyeball Test

To detect heteroscedasticity the residuals can be plotted against each measurement of the covariates in a scatter plot, which is done in figure 2.1. If the residuals do not plot well against a line, as in figure 2.1 (b), heteroscedasticity is present. This method of detecting heteroscedasticity is in *A Guide to Econometrics* referred to as the eyeball test. (Kennedy 2008, p. 116)

White's Robust Estimate

If the residuals appears to differ in variance, which would indicate heteroscedasticity, it is required to examine this issue further. This may be done by using White's robust estimate for the covariance matrix. This estimator can mathematically be described as equation 2.5, where $D(\hat{\mathbf{e}}^2)$ is a $n \times n$ diagonal matrix. (Lang 2013, p. 34)

$$\begin{aligned}\hat{C}ov(\hat{\beta}) &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t D(\hat{\mathbf{e}}^2)\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\hat{e}_i^2 x_i^t x_i\right)(\mathbf{X}^t\mathbf{X})^{-1}\end{aligned} \tag{2.5}$$

This makes it possible to compare the coefficient covariance from the ordinary least square (OLS) regression with the coefficient matrix incorporated in White's robust estimate. If heteroscedasticity is not present, these matrices will equal each other.

2.4.2 Solutions to Heteroscedasticity

The solution to this type of heteroscedasticity is to incorporate White's robust estimate in the regression. If heteroscedasticity is present this will improve the regression and it is therefore advisable to always incorporate White's robust estimate in the regression. (Lang 2013, p. 34)

2.5 Multicollinearity

Multicollinearity is a phenomenon where two or more of the covariates are related to each other in such a way that the quantitative measure of the variables are linearly dependent to a large extent. If some covariates are collinear, the ordinary least square (OLS) estimates of these parameters will have a large variance. (Kennedy 2008, p. 193)

A consequence of having a large variance is that the estimates are not precise and will therefore not work for hypothesis testing. When the OLS is used for prediction, multicollinearity will not be an issue. Another problem arise when trying to interpret a collinear relationship and not knowing what parameters influence one another. This may lead to specification errors. (Kennedy 2008, p. 194)

2.5.1 Detecting Multicollinearity

Detecting collinearity of two covariates can be done in different ways. Below are three common ways to examine the phenomenon.

Scatter Plot

Detecting multicollinearity can be done by putting all the measurements of each covariate in two separate, ordered, vectors and plotting them against each other. This way a scatter plot is constructed and if multicollinearity exists the measurements should be scattered around a straight line as described by the example in figure 2.2

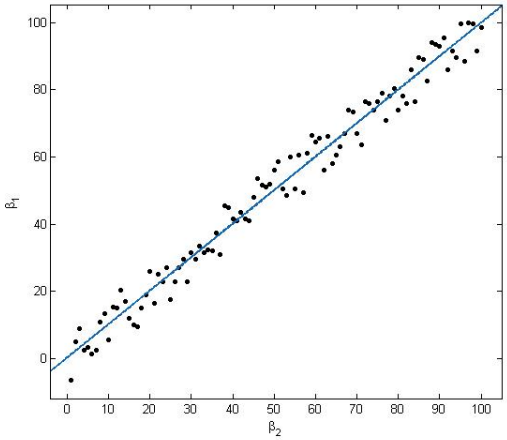


Figure 2.2: Positive multicollinearity between β_1 and β_2.

Correlation Matrix

A second way to detect multicollinearity is to construct the correlation matrix (Kennedy 2008, p. 195):

$$\mathbf{R}(\mathbf{X_1}, \mathbf{X_2}) = \frac{\mathbf{Cov}(\mathbf{X_1}, \mathbf{X_2})}{\sqrt{\mathbf{Cov}(\mathbf{X_1}, \mathbf{X_1})\mathbf{Cov}(\mathbf{X_2}, \mathbf{X_2})}} \tag{2.6}$$

The off-diagonal elements in $\mathbf{R}$ represents the correlation coefficients for the data in question. A correlation coefficient above 0.8 indicates a high correlation between the variables.

Variance Inflation Factor, VIF

A third way to detect multicollinearity is to calculate the VIF-value for each covariate in the model. The VIF-value can be expressed mathematically as

$$VIF = \frac{1}{(1 - R^2)} \tag{2.7}$$

The R^2 value is here represented by the R^2 achieved when doing a regression for each individual covariate with the rest of the covariates as independent variables. Hence, each comparison of two potential collinear variables has its specific VIF-value. The VIF-values can also be achieved by taking the inverse of the correlation matrix in section 2.5.1. If a VIF-value is greater than 10, there is an indication of harmful multicollinearity in the data. (Kennedy 2008, p. 199)

2.5.2 Solutions to Multicollinearity

Solutions to multicollinearity varies and there is no definite solution that applies to all situations. The problem at hand is to reduce the variance of the estimated covariates.

One solution is to obtain more data. Another solutions is to omit one of the collinear variables, a problem that arises then is that the estimates of the remaining variables will be biased. For dummy variables, as mention in section 2.1.1, multicollinearity is conveniently solved by using a benchmark.

A Guide to Econometrics suggests for two *rules of thumb* when dealing with multicollinearity (Kennedy 2008, pp. 194-197):

1. "Don't worry about multicollinearity if the R^2 from the regression exceeds the R^2 of any of the independent variable regressed on the other independent variables."

2. "Don't worry about multicollinearity if the t-statistics are all greater than 2."

2.6 Model Validation

When using regression in order to create a predictive model it is important to examine how well the model represents the data it is derived from and to what extent it is possible to use the model for predictive purpose. This type of analysis is referred to as model validation and may be done with different types of statistical tools. In this section we present the tools that we will later use in chapter 4 of this thesis.

2.6.1 R^2 and Adjusted R^2

R^2 is a *measure of goodness of fit*. It measures how well the covariates in the model explains the variance in the dependent variable. R^2 is equal to the square of the sample correlation coefficient between y and $x\hat{\beta}$ (Lang 2013, p. 23):

$$R^2 = \frac{Var(x\hat{\beta})}{Var(y)} \tag{2.8}$$

The sample variance of y can be decomposed into two terms:

$$Var(y) = Var(x\hat{\beta}) + Var(\hat{e})$$

Thus, R^2 can also be expressed as

$$R^2 = 1 - \frac{Var(\hat{e})}{Var(y)} \tag{2.9}$$

It follows from equation 2.9 that the model should have as high R^2 as possible since this minimizes the error term $\hat{e}$ and therefore implies an improved estimation of the dependent variable y. (Lang 2013, p. 23) This would lead to the choice of a relationship with too

many covariates in it, since the addition of a covariate cannot cause the R^2 statistic to fall. (Kennedy 2008, p. 79)

The adjusted R^2, often denoted $\bar{R}^2$, solves this problem by adjusting for the degrees of freedom. (Kennedy 2008, p. 79) This implies that $\bar{R}^2$ could fall if an additional covariate accounts for only a small amount of the unexplained variation in the dependent variable, where R^2 definitely increases. An extra covariate should therefore only be seriously considered for inclusion in the set of covariates if $\bar{R}^2$ rises. This suggests that econometricans should search for the optimal set of covariates by determining which set of covariates produces the highest $\bar{R}^2$. (Kennedy 2008, p. 80)

2.6.2 Hypothesis Testing

Hypothesis testing is a method of using statistics in determining the probability that a hypothesis is true. The process of hypothesis testing usually consists of four steps (MathWorld 2014):

1. The first step is to formulate a null hypothesis H_0 and an alternative hypothesis H_a. The null hypothesis implies that the observations are a result of pure chance, while in the alternative hypothesis the outcome of the observations are caused by a pattern or the distribution under question.

2. The second step is to identify a test statistic that can be used to access if the null hypothesis is true.

3. The third step is to calculate the p-value. When assuming that the null hypothesis is true: the p-value is the probability that the test statistic is at least as significant as the one observed. A smaller p-value means stronger evidence against the null hypothesis.

4. The fourth step is to compare the p-value with a significant value α. If $p \leq \alpha$ the null hypothesis is ruled out and the alternative hypothesis is accepted; i.e., the observed effect is statistically significant.

2.6.3 F-statistic and p-value

In regression it is also common to report a p-value for each covariate. This p-value is achieved by first calculating the F-statistic. When the error terms are normally distributed the F-statistic is calculated as (Richard A. DeFusco et. al. 2007)

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}} \tag{2.10}$$

In equation 2.10: y_i are the observed values, $\hat{y}_i$ the estimated values[1] and $\bar{y}$ the average of the dependent variable. n is the number of observations and k is the number of covariates in the regression. The F-statistic follows the F-distribution, which can be viewed in figure 2.3.

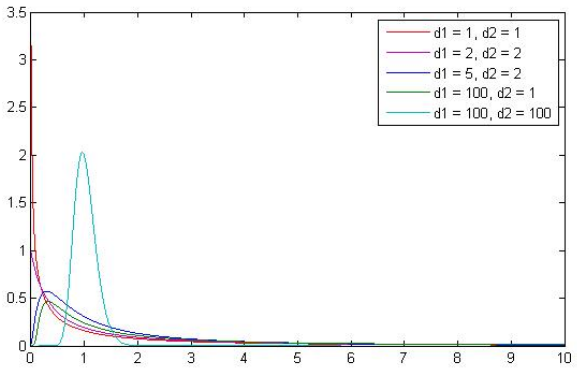


Figure 2.3: F-distribution, where $d1$ = degrees of freedom in the numerator and $d2$ = degrees of freedom in the denominator.

The p-value can then be achieved by comparing the F-statistic with the area under the specific distribution. This implies that a large F-statistic leads to a small p-value. When the p-value is used for hypothesis testing, the null hypothesis is that the corresponding coefficient is equal to zero. This may be illustrated mathematically as

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

In regression it is also common to report a F-statistic for the hypothesis that all covariates are equal to zero. This value is computed as (Lang 2013, p. 24)

$$F = \frac{n-k-1}{k} \frac{R^2}{1-R^2} \tag{2.11}$$

This is interpreted in the same way as the F-statistic for the individual covariates and is usually the first value to consider when checking the regression. Thereafter, the t-test is used for checking the significance for each individual covariate.

[1]The definition of the estimated values of y_i: $\hat{y_i} = \hat{\beta_0} + x_{i1}\hat{\beta_1} + \cdots + x_{ik}\hat{\beta_k} \qquad i = 1, 2, \ldots, n$

2.6.4 t-test

A t-test is a common way of hypothesis testing. In regression it is used to confirm whether the covariates x_i are significant. More precisely the aim in backward elimination is to find a model where all covariates are significant. Hence, the t-test is done for each individual covariate.

In a t-test, the null hypothesis, H_0, is that the covariate $x_{\bullet j}$ is not explanatory for the dependant variable and thus the respective coefficient β_j is zero. The alternative hypothesis, H_a is that the coefficient explains a part of the dependant variable and thus β_j is not zero.

This may be illustrated mathematically as

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

In a t-test the test statistic is computed for each $\hat{\beta}_j$ as

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{2.12}$$

The t-value for the estimated β_j is then compared to the t-distribution and if the t-value falls in the region specified by the selected confidence level, the null hypothesis is regarded as supported.

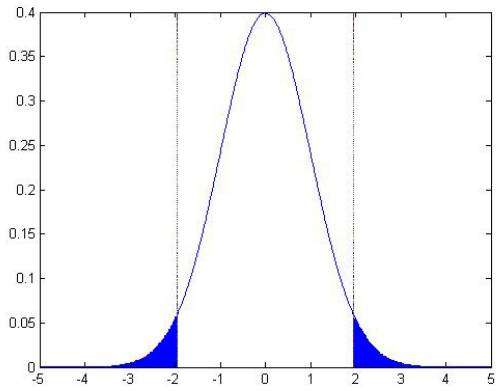


Figure 2.4: The t-distribution with $n = 10,000$ and a confidence interval of 95%, represented by the white area below the graph.

When a regression is done and estimates of the covariates are found, each specific t-value can be compered to see if they are within the confidence interval; represented in figure 2.4 by the white area below the graph.

2.6.5 Residual Analysis

The second assumption of the multiple linear regression model states that the expected value of the error term is zero. This is however seldom the case in practical applications. It is thus of importance to study the residual in order to examine in what extent assumption two may be violated. This will make it possible to recognise patterns in the residual that could increase the understanding of the regression and eventually improve it. This is referred to as residual analysis.

We recall the regression equation:

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + e_i \qquad i = 1, 2, \ldots, n \tag{2.13}$$

When the regression is done and estimates of β_j are determined, the residuals $\hat{e_i}$ can be achieved by the following manipulation of equation 2.13 (Lang 2013, p. 26):

$$\hat{e_i} = y_i - \left(\hat{\beta}_0 + x_{i1}\hat{\beta}_1 + \cdots + x_{ik}\hat{\beta}_k\right) \qquad i = 1, 2, \ldots, n \tag{2.14}$$

Histogram

The residuals can be illustrated in a histogram, as in figure 2.5. If the residuals are normally distributed around zero, the second assumption is regarded as valid.

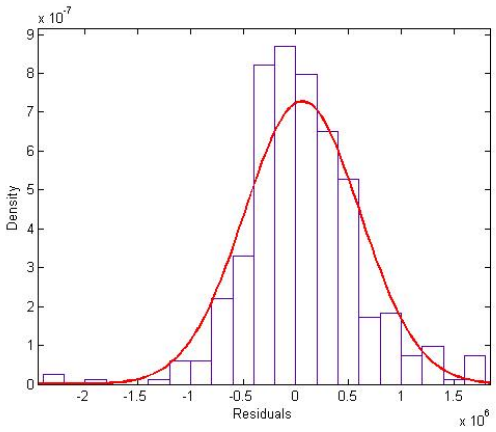


Figure 2.5: Histogram of residuals.

Normal Probability Plot

A normal probability plot is another way of displaying the residuals in order to see if they are normally distributed. If the probability plot follows a straight line the residuals are normally distributed and thus the second assumption is regarded as valid. (Richard M. Heiberger 2004, p. 110) The normal probability plot is constructed by arranging the residuals from the

smallest to the largest, and plotting them against the theoretical values they would have if they are normally distributed; i.e.,

- Vertical axis: Ordered response values.
- Horizontal axis: Normal order statistic medians or means.

2.6.6 Cross-validation

An issue that occurs when validating a regression model using residual analysis is that the model is made from the same data that is used for testing the model with residual analysis. The problem may be described as the model knows the data too well. In *The Collected Works of John W. Tukey - Philosphy and Principles of Data Analysis* it is described as (Tukey 1986):

> "...the procedure will likely work better for these data than for almost any other data that will arise in practice. The apparent degree of fit will almost never be representative..."

A solution to this issue is cross-validation, which can be done in different ways but with the common purpose of testing the model on data that has not been used for deriving the model. (Tukey 1986, p. 638)

One way of using cross-validation is to randomly select and remove a part of the data before the regression is performed. For example, remove 5% of the data and when the regression on the remaining 95% of the data is performed and $\hat{\beta}_j$ are determined; the estimates of y_i can be computed as

$$\hat{y_i} = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + \cdots + x_{ik}\hat{\beta}_k \qquad i = 1, 2, \ldots, n \tag{2.15}$$

This is done for the 5% of the data that was removed and estimates of these y_i values are computed. These estimates are compared with the real y_i values and if they coincide to a large extent the regression can be viewed as an accurate regression. (Tukey 1986)

3 Method

The method of this thesis is focused on narrowing down large amounts of data in order to see relevant patterns and relationships between variables. The data for this thesis is achieved from the web-service Slutpris. (Sigot 2014) It consists of 11,006 observations of the following variables: *address, area, balcony, construction year, elevator, fireplace, floor number, maisonette, monthly fee, penthouse, postal code, price, reserve price, rooms and sales date*. The observations are mainly from the Stockholm City Centre during the period 2012-01-01 to 2013-12-27.

The mathematical method used to determine the results in this thesis is the multiple regression model and as described in section 2.2 there are certain assumptions that have to be met in order for the regression to be valid. These assumptions are examined with relevant statistical tools.

The method can be described as a three step method shown below and the rest of the method is ordered in the same way:

1. Data is collected and preprocessed.
2. The model for predicting apartment prices in Stockholm City Centre is determined and interpreted.
3. Assumptions for the regression to be valid are examined.

3.1 Data Pre-processing

When preparing the data the postal codes starting with the following numbers have been removed: *100, 101, 102, 104, 120, 121, 123, 126, 128, 130, 135, 167, 168 and 171*. This was a total of 414 observations and left where observations with the postal codes: *111, 112, 113, 114, 115, 116, 117 and 118*. The observations were removed since they did not have valid postal codes for the Stockholm City Centre. Moreover, in the following order, 17 observations with unknown area, 59 with unknown floor number, 28 with unknown monthly fee and 2,783 that have an unknown construction year were removed. Hence, left are 8,164 observations.

3.1.1 Adjusting the Price Variable

The *price* variable was adjusted to reflect the monthly change in price during the period 2012-01-01 to 2013-12-27. The HOX Stockholm BR Index was used to adjust for these price differences. It is an index developed by Valueguard in conjunction with KTH Royal Institute of Technology. The index is based on data from, among others, Mäklarstatistik AB, which compiles data from the Swedish real estate agents. It gives an adequate picture of price trends for apartments in Stockholm. (Valueguard 2014) When adjusting the *price* variable the date 2013-12 was used as a reference. The index values for the current period can be found in table 3.1.

Table 3.1: Index over price differences during the sales period

2012	**HOX Stockholm BR Index**	**2013**	**HOX Stockholm BR Index**
jan	170.24	jan	183.72
feb	172.07	feb	186.21
mar	175.15	mar	188.75
apr	174.87	apr	190.09
may	176.56	may	190.74
jun	174.94	jun	193.32
jul	178.23	jul	194.63
aug	178.70	aug	197.56
sep	178.88	sep	199.24
oct	179.63	oct	199.73
nov	178.50	nov	202.24
dec	178.74	dec	202.79

3.2 Variable Selection

When the optimal model was decided it was of essence to have a presentiment of what factors that are of importance when valuing an apartment in the Stockholm's City Centre. Initially all factors that we believed could have an impact on the price were included in the model and a step-by-step process was performed in order to reduce and simplify the model to only include the variables that have a statistically significant impact on the price. This step-by-step process helped minimize the possibility that a variable or combination of variables, which could be of essence for the model, were disregarded or overlooked.

A rule that was used to determine the optimal model was to not include any covariate or dummy variable that had a p-value of 5% or more; i.e., avoid taking a risk of 5% or more when implying that a specific variable has a statistically significant impact on the price of an apartment. In addition to this, covariates that did not contribute to a higher $\bar{R}^2$ were excluded from the model.

3.2.1 Excluded Variables

The covariates *maisonette* and *number of rooms* were excluded from the model. Table 3.2 contains information about these variables as well as the reason for excluding them.

Table 3.2: Variables excluded from the model

Variable	**Unit**	**Comment**
Maisonette	Dummy	States if the apartment has a maisonette (Swedish: etage). The variable was excluded since it had a p-value of 20% and did not contribute to a higher $\bar{R}^2$.
Number of rooms: 1	Dummy	States if the apartment is a studio.
Number of rooms: 2	Dummy	States if the apartment is two-room apartment.
Number of rooms: 3	Dummy	States if the apartment has three rooms.
Number of rooms: 4	Dummy	States if the apartment has four rooms.
Number of rooms: 5 or more	Benchmark	States if the apartment has five or more rooms. The number of rooms variables were excluded since they did not contribute to a higher $\bar{R}^2$.

3.3 The Final Model

The final model used in this thesis will use the covariates *area and monthly fee*; dummy variables *balcony, construction year: 1336-1919, construction year: 1920-1959, construction year: 1960-1999, district: Gärdet, district: Kungsholmen, district: Norrmalm/Gamla stan, district: Södermalm, district: Vasastan, elevator, fireplace, ground floor and penthouse* and benchmarks *construction year 2000-2013 and district: Östermalm* to predict the dependent variable *price*. Table 3.3 contains information about the variables used in the model.

Table 3.3: Variables in the model

Variable	**Unit**	**Comment**
Area	m^2	States the total area of the apartment.
Balcony	Dummy	States if the apartment has a balcony. French balconies are not included in this variable.
Construction year: 1336-1919	Dummy	States if the building the apartment is located in is constructed between 1336-1919. Further information in subsection 3.3.1.
Construction year: 1920-1949	Dummy	States if the building the apartment is located in is constructed between 1920-1949. Further information in subsection 3.3.1.
Construction year: 1950-1999	Dummy	States if the building the apartment is located in is constructed between 1950-1999. Further information in subsection 3.3.1.
Construction year: 2000-2013	Benchmark	States if the building the apartment is located in is constructed between 2000-2013. Further information in subsection 3.3.1.
District: Gärdet	Dummy	States if the apartment is located in the area we have defined as Gärdet. Further information in subsection 3.3.2.
District: Kungsholmen	Dummy	States if the apartment is located in the area we have defined as Kungsholmen. Further information in subsection 3.3.2.
District: Norrmalm/Gamla stan	Dummy	States if the apartment is located in the area we have defined as Norrmalm/Gamla stan. Further information in subsection 3.3.2.
District: Södermalm	Dummy	States if the apartment is located in the area we have defined as Södermalm. Further information in subsection 3.3.2.
District: Vasastan	Dummy	States if the apartment is located in the area we have defined as Vasastan. Further information in subsection 3.3.2.
District: Östermalm	Benchmark	States if the apartment is located in the area we have defined as Östermalm. Further information in subsection 3.3.2.
Elevator	Dummy	States if there is an elevator in the building where the apartment is located.
Fireplace	Dummy	States if the apartment has a fireplace.
Ground floor	Dummy	States if the apartment is located at the ground floor; i.e., floor number 0 or 0.5.
Monthly fee	SEK	States the monthly fee for the apartment.
Penthouse	Dummy	States if the apartment is located at the top floor. This dummy is also associated with attributes such as windows in several directions, visible ceiling struts, good view and the subjective value of owning a penthouse.
Price	SEK	States the price that the apartment was sold for.

3.3.1 Dummy Variables for Construction Year

The construction year of the building, in which the apartment is located, will have an impact on the price of the apartment. Different time spans characterises various qualities such as ceiling height, ground plans and atmosphere. The data used in this thesis includes information about the construction year for the different observations, which ranges from year 1336 to 2013. The classification of the various time spans is based on our experience and from information acquired from Hemnet. (Hemnet 2014)

The first time span is set from year 1336 to 1919. These apartments are characterised by a ceiling height of over three meters as well as a unique coveted atmosphere. However, they often lack well laid out ground plans. Next time span is set between the years 1920 and 1949. The apartments in this time span typically have 2.7 to 3.0 meters in ceiling height and good ground plans. The third time span, which is set between the years 1950 and 1999, is often characterised by apartments with a low ceiling height of 2.3 to 2.5 meters and an austere atmosphere. Their ground plans are however, most often, very efficient. The last time span is set from year 2000 to 2013. These apartments usually have a ceiling height of 2.5 to 2.7 meters, a modern design and very good ground plans.

3.3.2 Dummy Variables for District

The data used in this thesis includes information about the addresses and postal codes for the various observations. With this information it is possible for the observations to be positioned within their respective districts. The districts were selected with regards to their locations and differences in price. The partition of the districts was done with the help of Google Maps (Maps 2014), Hemnet (Hemnet 2014) and Svensk Mäklarstatistik AB (Mäklarstatistik 2014). The different districts and their locations are shown in figure 3.1.

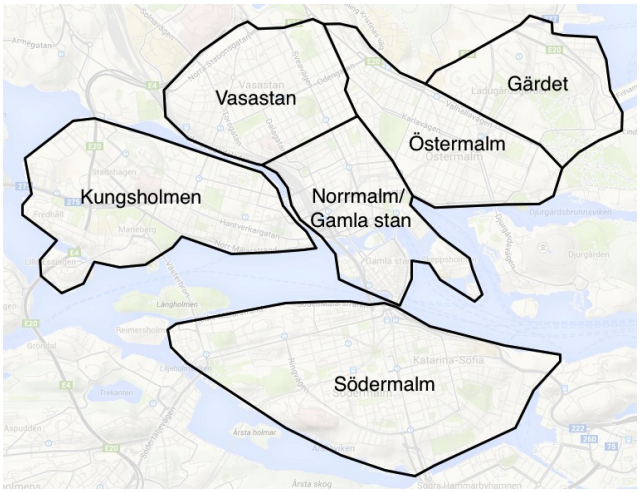


Figure 3.1: Districts of Stockholm City Centre, represented by dummy variables.

3.4 Model Checking

In this section we examine the assumptions for the multiple regression model mentioned in section 2.2. Relevant statistical tools described in the background is used for the examinations and issues are dealt with when necessary. All five assumptions are regarded as confirmed and therefore the regression performed on the data is regarded as valid.

3.4.1 Assumption 1: Linearity Between Covariates and the Dependent Variable

This assumption is examined for each covariate that is not a dummy variable, thus only the variables *area* and *monthly fee*. This is done by constructing a scatter plot for these covariates against the dependent variable *price* and fitting a line to these points. As described below, both *area* and *monthly fee* show a linear relationship with *price* and therefore this assumption is regarded as confirmed. The rest of the variables are dummy variables and will thus not be examined.

Area

A regression with the single covarite *area* explains approximately 80% of the dependent variable *price* and it is therefore the most important variable to examine. We start with this variable and construct the scatter plot in figure 3.2. As it can be seen, there is a linear relationship between *area* and *price*.

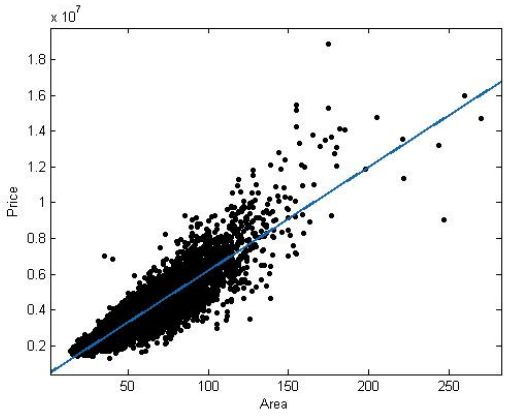


Figure 3.2: Linear relationship between *price* and *area* with $R^2 = 0.81$.

Monthly Fee

The variable *monthly fee* does not show an as clear linear relationship with the dependent variable *price* as the variable *area*. Although, as can be viewed in figure 3.3, there is an indication of a linear relationship between *monthly fee* and *price*.

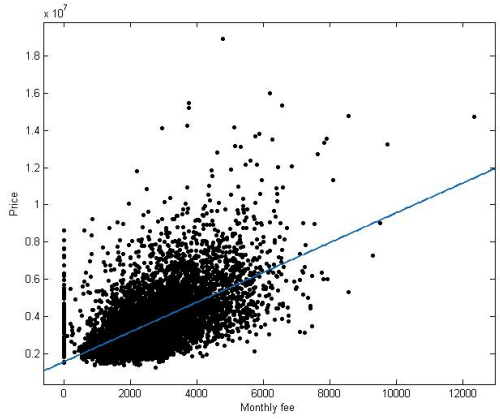


Figure 3.3: Linear relationship between *price* and *monthly fee* with $R^2 = 0.34$.

3.4.2 Assumption 2: Expected Value of the Error Term is Zero

This assumption is evaluated using cross-validation, described in subsection 2.6.6, with 5% of the data used for the cross-validation. Thus, 5% of the data was randomly selected and removed before the regression was done. Then, using the coefficients from the output of the regression – displayed in table 4.1 – prices for these apartments were modelled. These modelled prices are called $ModelPrice$ and were compared with the real prices – $Price$. This is shown in the figure 3.4, where a 95% confidence interval is included in the graph.

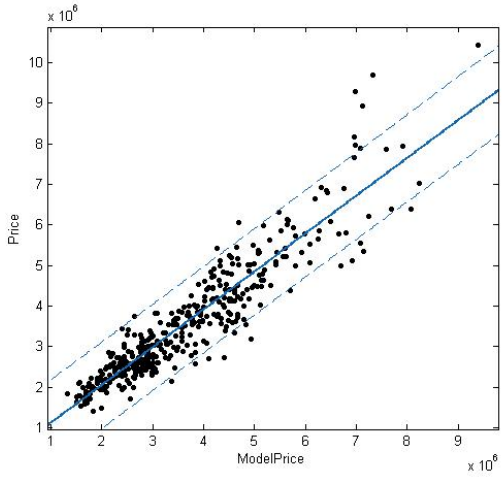


Figure 3.4: Linear relationship between the price from to the model and the real price with a 95% confidence interval.

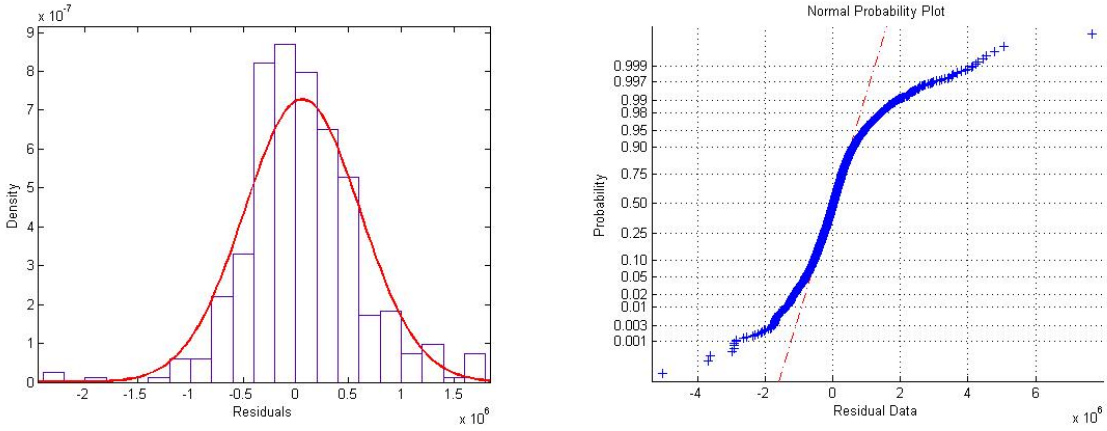
If our model would be able to perfectly predict the price of an apartment, the *modelled prices* would equal the *real prices* and the scatter plot in figure 3.4 would follow a straight

line and as suggested by this assumption *the expected value of the error term would be zero.* This is not the case for each individual observation, but as seen in figure 3.4 most of the *modelled prices* are within a confidence interval of 95%.

To examine this assumption further, the residuals of each comparison can be displayed in a histogram and in a normal probability plot – see figure 3.5 – where it is clear that the residuals are normally distributed around zero. Thus, this assumption is regarded as confirmed.

The residuals are calculated by subtracting $ModelPrice$ from $Price$:

$$Residual = Price - ModelPrice \tag{3.1}$$



(a) Histogram of the residuals.

(b) Normal probability plot of the residuals.

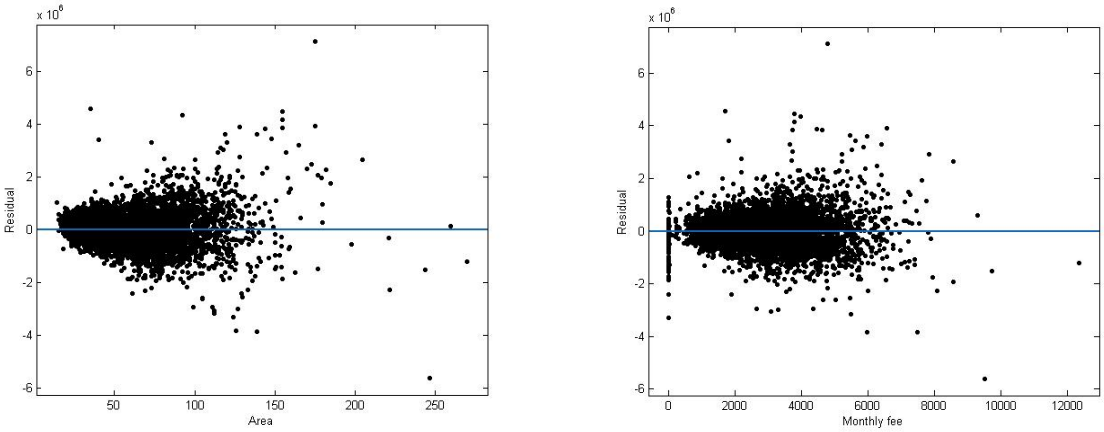
Figure 3.5: Residual analysis.

3.4.3 Assumption 3: Homoscedacticity

In accordance with the background, assumption three is evaluated by examining if homoscedacticity is present in the model.

Homoscedacticity

Homoscedacticity is examined for the covariates *area* and *monthly fee* by creating two scatter plots of these variables and the residuals in the regression. These plots are shown in figure 3.6.



(a) Area

(b) Monthly fee

Figure 3.6: Hetroscedasticity among the covariates *area* and *monthly fee*.

For both of these variables it is not definite whether the residuals are hetroscedastic or not. It is therefore necessary to further examine this by incorporating White's robust estimate in the regression. By comparing the two subsequent coefficient covariance matrices it can be seen that these matrices does not equal each other and thus we should incorporate White's robust estimate in the regression. The coefficient covariance matrices are displayed in table A.1 and table A.2 in the appendices.

The assumption of homoscedasticity is a difficult issue, but by using White's robust estimate we argue that we have taken this assumption in consideration and thus can say that this assumption is regarded as confirmed.

3.4.4 Assumption 4: Measurement Errors

The fourth assumption, which states that the covariates can be considered fixed in repeated samples, is evaluated by assessing what effects autoregression and errors in variables will have on the model.

Autoregression should not be an issue in the model since we are not using any covariates that could be a lagged value of the dependent variable. The covariate *area* has a different unit and *monthly fee* has a lot smaller magnitude compered to *price*.

Errors in variables represent a greater threat to the reliability of the model. The data used in this thesis is achieved from the web-service Slutpris. It is an independent organisation that provides sales data for apartments to the public. The data are based on information that brokers publish in their sales-prospects. Slutpris states that they cannot guarantee that the data for every object is correct. (Sigot 2014) However, based on reputation that the web-service Slutpris has it is expected in this thesis that the vast majority of the objects have the correct data and the possible errors, after the data preparation, should not have a

significant impact on the result. For more information regarding the data preparation see section 3.1.

3.4.5 Assumption 5: Multicollinearity

Multicollinearity among the covariates is examined for the variables *area* and *monthly fee*. The rest of the variables are dummy variables and multicollinearity among these variables is prevented by using a benchmark.

In accordance with section 2.5.1, multicollinearity between *area* and *monthly fee* is examined by first creating a scatter plot over these variables as shown in figure 3.7.

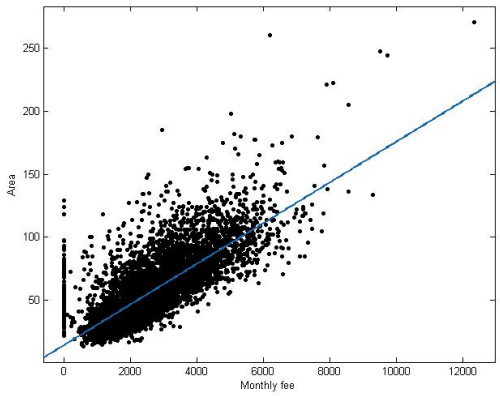


Figure 3.7: The covariates *area* and *monthly fee* plotted against each other with $R^2 = 0.584$.

We can see that there appears to be some relationship between the variables but the question is to what extent and if it should be considered as harmful to the interpretation of the model.

The correlation coefficients are

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.7607 \\ 0.7607 & 1.0000 \end{pmatrix} \tag{3.2}$$

As we can see the values of the correlation coefficient are less than 0.8, which implies that multiocollinearity is not a problem between *area* and *monthly fee*. The VIF-values are computed by taking the inverse of $\mathbf{R}$ with the result

$$\mathbf{R^{-1}} = \begin{pmatrix} 2.3735 & -1.8056 \\ -1.8056 & 2.3735 \end{pmatrix} \tag{3.3}$$

Since the VIF-value 2.3735 is less than 10, it is not considered to be any harmful multicollinearity present in the data and this assumption is therefore regarded as confirmed.

4 Result

The regression is performed in MATLAB using the function `LinearModel.fit`. The result of the regression is displayed in table 4.1.

Table 4.1: Regression output with White's robust estimators

Variable	**Estimate**	**SE**	**tStat**	***p*-value**	**95% confidence interval**	
Intercept	642 060.00	37 448.00	17.15	1.06E-64	568 650.00	715 470.00
Area	60 825.00	342.14	177.78	0	60 150.00	61 500.00
Balcony	125 570.00	11 223.00	11.19	7.73E-29	103 570.00	147 570.00
Construction year: 1336-1919	482 500.00	26 339.00	18.32	2.06E-73	430 870.00	534 140.00
Construction year: 1920-1949	205 670.00	24 130.00	8.52	1.85E-17	158 370.00	252 970.00
Construction year: 1950-1999	-69 650.00	25 644.00	-2.72	6.62E-03	-119 920.00	-19 380.00
District: Norrmalm/Gamla stan	-134 410.00	36 089.00	-3.72	1.97E-04	-205 150.00	-63 660.00
District: Kungsholmen	-370 550.00	22 078.00	-16.78	3.98E-62	-413 830.00	-327 270.00
District: Vasastan	-211 180.00	21 959.00	-9.62	8.95E-22	-254 230.00	-168 140.00
District: Gärdet	-278 490.00	25 578.00	-10.89	2.08E-27	-328 630.00	-228 350.00
District: Södermalm	-417 670.00	21 185.00	-19.72	1.82E-84	-459 200.00	-376 150.00
Elevator	155 300.00	14 048.00	11.06	3.36E-28	127 770.00	182 840.00
Fireplace	192 340.00	16 571.00	11.61	6.83E-31	159 860.00	224 820.00
Ground floor	-237 800.00	13 409.00	-17.74	5.17E-69	-264 090.00	-211 520.00
Monthly fee	-210.51	6.90	-30.52	2.62E-193	-220.00	-200.00
Penthouse	326 420.00	29 973.00	10.89	2.02E-27	267 660.00	385 170.00
Number of observations:	7 756.000					
Error degrees of freedom:	7 740.000					
Root mean squared error:	450 000.000					
R-squared:	0.914					
Adjusted R-squared:	0.914					
F-statistic vs. constant model:	5 520.000					
p-value:	0.000					

4.1 Model Validation

Model validation is first done using residual analysis and cross-validation. Secondly the R^2 is evaluated and lastly the t-statistics and p-values for the covariates are evaluated.

4.1.1 Residual Analysis and Cross-validation

The regression is performed without incorporating for White's robust estimators and a residual analysis on this result is done with a histogram over the residuals, which can be viewed in figure 4.1. A cross-validation on this result can be seen in figure 4.2.

Next, the regression is performed with incorporating for Whites's robust estimators and a residual analysis is performed on the result with a histogram over the residuals. This can be viewed in figure 4.1 and a normal probability plot over the Whites's robust resudials can be seen in figure 4.3. The result is also evaluated by cross-validation that can be viewed in figure 4.2.

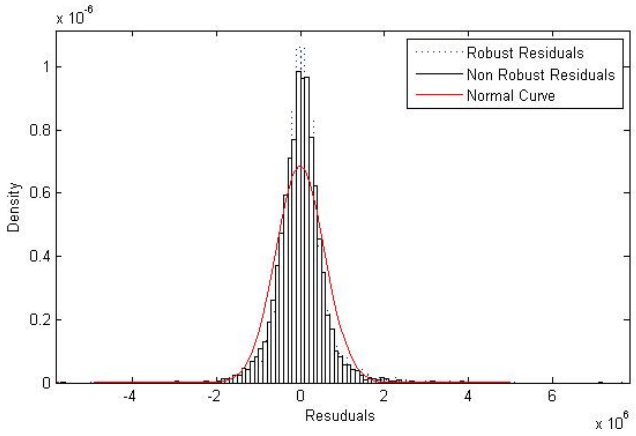
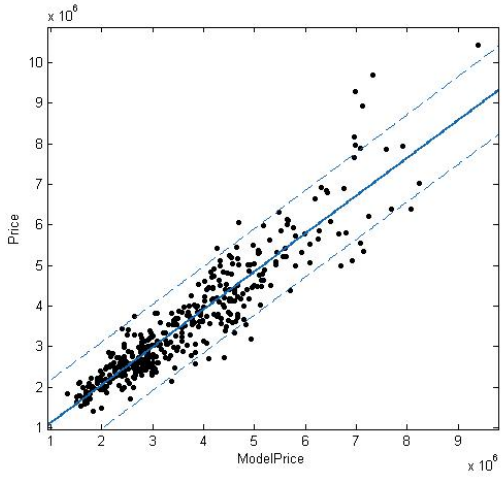
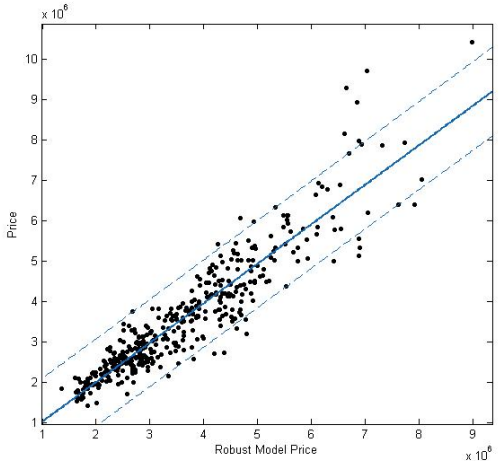


Figure 4.1: Histogram over the regression, with and without, Whites's robust estimates.



(a) Modelled Prices



(b) Modelled prices with Whites's robust estimates

Figure 4.2: Cross-validation on 5% of the original data.

Figure 4.1 and figure 4.2 indicates that there is not a major difference between the regression with or without White's robust estimates. However, in the final model we have used the White's robust estimates anyway since it is advisable to always incorporate it in the regression. (Lang 2013, p. 34)

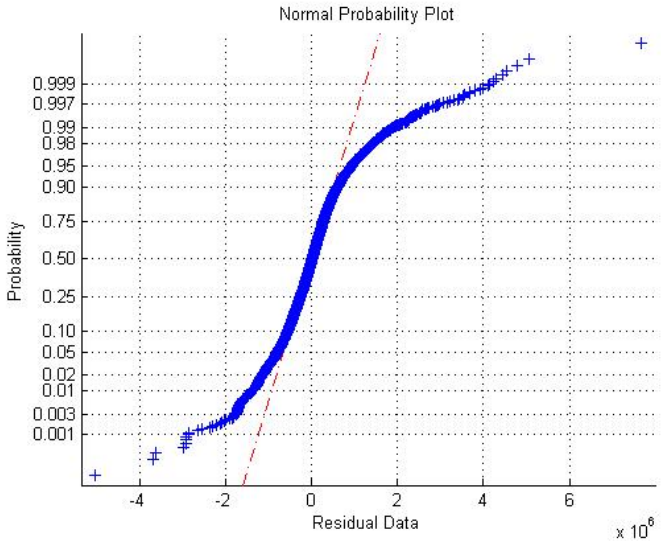


Figure 4.3: Normal probability plot over the Whites's robust residuals.

4.1.2 Evaluating the R^2, t-statistics and p-values

The R^2 is 0.91 for the regression, which shows that the model explains 91% of the variation in the data. All the t-statistic values are less than -2 or greater than 2. This supports

the alternative hypothesis that each covariate are significant, with a confidence interval of 95%. Furthermore, the p-values are low (≤ 0.05) and this also supports the the alternative hypothesis that each covariate are significant.

4.2 Regression Equation

We recall the output of the regression in table 4.1. This result can be illustrated in accordance with the regression equation and the estimated value of price:

$$\begin{aligned} Price = & 642060 \\ & + 60825 \times Area \\ & + 125570 \times Balcony_{0,1} \\ & + ConstructionYear_{0,1,2,3} \\ & + District_{0,1,2,3,4,5} \\ & + 155300 \times Elevator_{0,1} \\ & + 192340 \times Fireplace_{0,1} \\ & - 237800 \times GroundFloor_{0,1} \\ & - 210 \times MonthlyFee \\ & + 321420 \times Penthouse_{0,1} \end{aligned} \tag{4.1}$$

The dummy variables have an index showing that they can take on either the values 0 or 1. $ConstructionYear$ and $District$ can take on four respective six values and these values are displayed in table 4.2. Using this equation there is a 91% explanation degree of the *price*.

Table 4.2: Values for the dummy variables $ConstructionYear$ and $District$

	ConstructionYear	
0	Construction year: 1999-2013	0
1	Construction year: 1336-1919	482 500.00
2	Construction year: 1920-1949	205 670.00
3	Construction year: 1950-1999	-69 650.00
	District	
0	District: Östermalm	0.00
1	District: Norrmalm/Gamla stan	-134 410.00
2	District: Kungsholmen	-370 550.00
3	District: Vasastan	-211 180.00
4	District: Gärdet	-278 490.00
5	District: Södermalm	-417 670.00

5 Discussion

A perfect model that is able to predict the exact value of an apartment is very hard to achieve – if not impossible – since it has to include all aspects of what makes an apartment valuable. Trying to achieve such a model would be very complex with variables that are difficult to measure and may differ among people. Therefore we believe in having a more simple model that is easy to understand and use, which at the same time predicts the value of an apartment reasonably well. This model will give a reasonably well depiction of how important different factors are when valuing an apartment in the Stockholm City Centre.

5.1 Indications of the Covariates

A discussion will follow in this section regarding what the values of the covariates indicates as well as possible error sources for the covariates. When discussing a covariate, the rest of the covariates in the model will be considered fixed.

Area

Area is the most statistically significant covariate for the regression and explains alone approximately 80% of the *price*. Table 4.1 shows that each square meter adds 60,825 SEK to the *price*.

Balcony

There are no information in the data concerning the appearance of the balcony, its size or in what directions it faces. Neither are there any information regarding if an apartment has more than one balcony. French balconies are not included in this covariate.

As shown in table 4.1, an apartment is worth 125,570 SEK more if it has a balcony. If a balcony faces freely the south in direction it is of course worth more, since it will be positioned for more sun hours. This is reasonable for mid-sized apartments. For larger apartments we belive it might be of inclination for the buyer to trade more than two square meters for a balcony; i.e., larger apartments with a balcony is worth more. The opposite might as well be true for small apartments.

Construction Year

As seen in table 4.1: The construction year of the building will have a significant impact on the *price* of the apartment. Apartments in older buildings are more expensive, except for the apartments in buildings constructed between the years 1950 and 1999 that are worth less than the benchmark.

The benchmark, apartments in buildings constructed between the years 2000 and 2013, according to the data from Slutpris, typically have a higher *monthly fee*. This could imply multicollinearity with *monthly fee* and have an impact on the result; i.e., the apartments in buildings constructed between the years 2000-2013 have a lower value in the model.

District

Table 4.1 shows that the district also have a significant impact on the *price*. The benchmark, which is Östermalm, is the most expensive. Followed by – in the following order – Norrmalm/Gamla stan, Vasastan, Kungsholmen, Gärdet and Södermalm.

Elevator

An elevator will add 155,300 SEK to the *price*, which can be seen in table 4.1. This is reasonable for apartments that are not situated at a floor near ground level. For these apartments, access to an elevator will probably not have as big of an impact on the *price*. The model however does not take this into account. The opposite is presumably true for apartments situated at a high floor.

Fireplace

There are no information in the data regarding the appearance of the fireplace or if the apartments have more than one fireplace. As shown in table 4.1: an apartment is worth 192,340 SEK more if it has a fireplace.

It is possible that this covariate could have multicollinearity with the covariate *construction year*, because it is more common that an apartment built between the years 1336 and 1919 has a fireplace. According to the data from Slutpris: 23% of the apartments built between the years 1336 and 1919 have a fireplace, while only 5% built between the years 2000 and 2013 have it. This could affect the covariate to have a higher value than it actually has.

Ground Floor

Apartments situated at the ground floor typically have poor view and insight from the street or the courtyard. These apartments will therefore have a lower *price*, which the model implies. As can be seen in table 4.1: the model states that apartments situated on the floor 0 or 0.5 are worth 237,800 SEK less.

Monthly Fee

The monthly fee determines the monthly cost for the apartment in consort with the interest of the loan on the apartment. For every 1 SEK more in *monthly fee* the apartment is worth 210.51 SEK less. What is included in the monthly fee will vary between the objects and will thus be a source of error in the model.

Penthouse

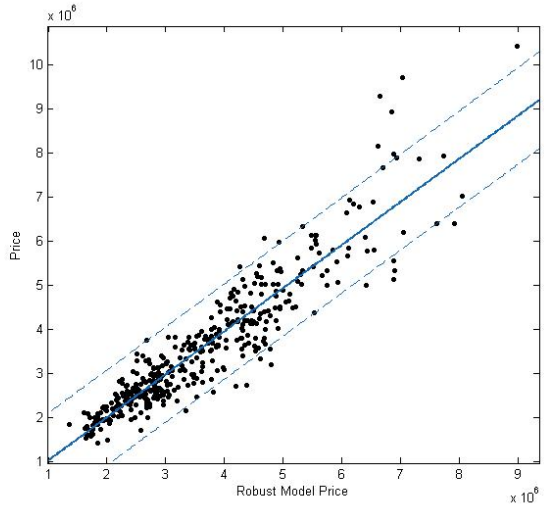
The covariate *penthouse* is not only associated with the apartment being situated topmost in the building. It is often connected with attributes such as windows in several directions, visible ceiling struts, good view and the subjective value of owning a penthouse. As shown in table 4.1: an apartment is worth 326,420 SEK more if it is a penthouse.

5.2 Conclusions

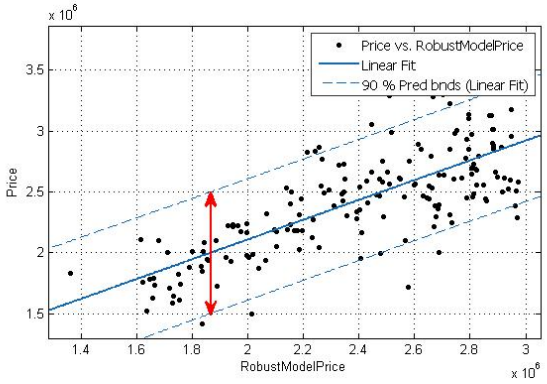
A conclusion that can be drawn from this thesis is that it is possible to predict the prices of apartments in Stockholm City Centre, using the covariates in table 3.3, with an explanation degree of 91%. The most important variable is the area of the apartment, which alone explains for 81% of the price. Second most important variable is the monthly fee explaining 34% of the price. The rest of the covariates in the model are dummy variables and influence the price of the apartments to various extents.

The cross validation with the confidence-interval for the predicted prices – presented in section 4.1.1 – can be interpreted as

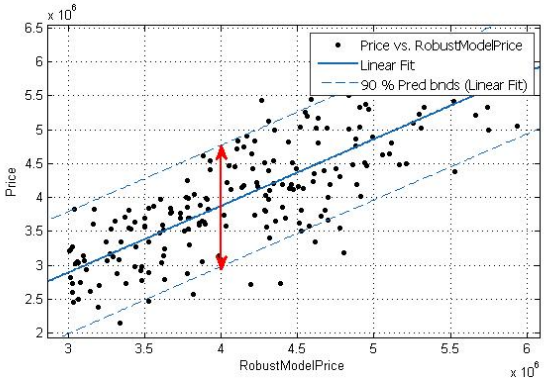
> *If you model an apartment in the Stockholm City Centre with the proposed model, your modelled price will fall in the price interval displayed in figure* 5.1(a).



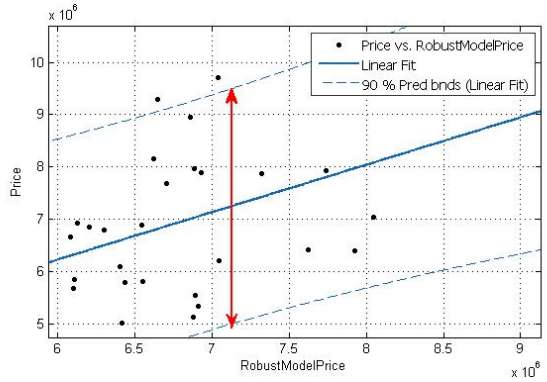
(a) All apartments, with a 95% confidence interval.



(b) Apartments in price range 0-3 million SEK



(c) Apartments in price range 3-6 million SEK



(d) Apartments in price range 6+ million SEK

Figure 5.1: Cross validation of the model divided in price ranges.

Figure 5.1(a) indicates that the 95% confidence interval is approximately two million SEK. Such a prediction interval makes the model appear useless for apartments that have a low value. However, since there are more observations in the data of apartments with a low value the model is actually more accurate for these apartments; e.i, the predictions are more closely clustered around a straight line. The opposite is as well true for higher priced apartments.

This makes it reasonable to divide figure 5.1(a) into subsections, where the confidence bounds differ between the subsections. In figure 5.1 this is done for illustrative purpose and three subsections are constructed, all with 90% confidence intervals. It appears that the prediction intervals differ a lot, from 1 million SEK in figure 5.1(b) to approximately 4.5 million SEK in figure 5.1(d). A conclusion can therefore be made that the model predicts lower priced apartments more accurately.

6 Proximity to Public Transport – a case-study and literature review

In this thesis we have been striving towards creating a model that can predict the price of an apartment. As mentioned, the perfect model would include all variables and aspects which one takes into account when valuing an apartment. Such a model would indeed be rather complex and due to difficulties of estimating various aspects and because of limited data one is forced to appreciate a less complex model which at the same time explains a good part of the apartment price.

One aspect which has not been included in the analysis so far, but which is suggested by Gunnvald and Gunnvald as an improvement possibility of a prediction model for apartment prices, is *proximity to public transport*. (R. Gunnvald and P. Gunnvald 2012) This aspect is evaluated by Strand and Vågnes (Strand and Vågnes 2001) for Oslo, Norway, with positive indicators of increased housing prices in relation to proximity to railroad and by Salon (Salon 2014) for Guangzhou, China, with the same result but for metro and Bus Rapid Transit Systems.

In this chapter we investigate the aspect of proximity to public transport in relation to apartment values. The main focus is to answer the question:

Is proximity to public transport an important aspect when valuing apartments?

6.1 Method

The methodology of this chapter is divided in two parts. At first, we try to answer the above question using a statistical approach in accordance with the theories about regression. This is done by evaluating data for sold apartments in the district Gärdet in Stockholm compared to their individual proximity to the closest subway station. Here we only see an indication of a correlation.

Secondly, we adapt a literary approach and take a step back in the analysis from looking at an individual level to looking at a society level of how public transport relates to creating sustainable suburban and urban areas in a city. This is done with theories about Large Technical Systems (Mayntz and Hughes 1988) and leads to the inevitable conclusion that

proximity to public transport is indeed an important aspect within city planning, but that the monetary aspects often need to be neglected because of the *technological trajectories* (Kuhn 1962) that determine the choice of the technical solutions.

The goal of the literary approach is to gain further understanding of how proximity to public transport relates to housing prices and further what determines how public transport is evolving as a technical system and what aspects that influence the decisions in the evolving process. Hence, articles on the subjects of public transport, city-planning and public transport in relation to housing and land value has been read. The articles has been found trough Google Scholar. Moreover, as we gradually have increased our knowledge about the development of public transport, relevant concepts and theories from innovation and management literature have been applied to further extend the analysis.

6.2 Statistical Analysis

For a variable to be regarded as explanatory for the dependent variable, in this case the sales price, there should be a seemingly linear relationship between the variables. A problem which arise when dealing with variables that only have a small impact on the predicted price is that they are hard to measure and that a very large sample is needed in order to be able to observe any patterns.

Proximity to public transport is indeed a measurable aspect that could be used in a model for apartment prices. The problem is that it is difficult to measure it. In this analysis we have limited the study to only look for patterns between proximity to subway station and the value of an apartment, but equally well one could also include for example proximity to buss stations. We have also limited our analysis to the district Gärdet in Stockholm. The choice of Gärdet is due to ongoing plans to invest in public transport in this area by Stockholms Läns Landsting. (Trafiknämnden 2011) This makes Gärdet an interesting case-study.

To investigate if there is a relationship between proximity to subway station and apartment prices at Gärdet we begin by displaying a scatter plot of the variables in question. Since the area of an apartment is very influential on the price of an apartment it is reasonable to only look for linear patterns between square meter price and proximity to subway station. This is illustrated in figure 6.1.

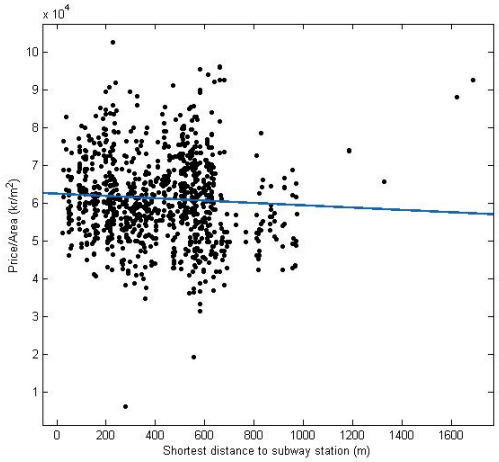


Figure 6.1: *Price/Area* (SEK/m^2) plotted against *distance to subway station* (m). Sample seize: 975 observations.

As can be viewed in figure 6.1, a linear relationship between proximity to subway station and square meter price cannot be found. However, as described, this relationship is hard to measure and we cannot form this sample say that there is not a relationship between the variables.

To further analyse how proximity to subway station can be used when trying to predict the apartment price a regression is performed with only *proximity to subway station* as covariate and *square meter price* as the dependent variable. The result of this regression is displayed in table 6.1.

Table 6.1: Regression between *square meter price* and *proximity to subway station*. Output from MATLAB

Variable	**Estimate**	**SE**	**tStat**	***p*-value**
Intercept	62521	775.92	80.57	0
Proximity to subway station	-3.07	1.64	-1.87	0.06

From table 6.1 it can be seen that the p-value for the hypothesis *proximity to subway station has no impact on the price*, hence the null hypothesis, is 0.06. This value is not as low as we would like it to be[1] but it still supports the decision to reject the null hypothesis in favour of the alternative hypothesis that *proximity to subway station has an impact on the price*.

From table 6.1 it can also be seen that the sign before the estimate of the covariate, proximity to subway station, is negative which supports the hypothesis that the square

[1] If the p-value is less than 0.05 it is regarded as statistically significant to reject the null hypothesis

meter price is lower the further away from a subway station the apartment is located. However, since the p-value is not as low as we would like it to be and because the R^2 value is low (< 0.05) one should not put too much confidence in this result.

6.3 Literature Review

When trying to evaluate apartment prices it is easy to rely too much importance on the individual characteristics of an apartment. However, as discussed in this thesis, understanding grows as more aspects are taken into account. One such aspect is proximity to public transport. Although this aspect clearly may be an aspect for the individual prospector, complexity arises when including opportunity costs of not having any nearby alternatives for public transport. Another aspect which complicates the matter for the individual prospector is related to the theory about *the tragedy of the commons* (Hardin 1968), which in this case can be interpreted as *public transport is not something that you can choose to pay for or not.* Public transport is rather a matter subject to decisions within the government and is thus financed by the taxpayers. An example of this is the government proposition of investing 522 billion SEK during the period of 2014 to 2025 in infrastructure projects in Sweden. (Kollektivtrafik 2012) In the same proposition, decisions for public transport solutions are also proposed and evaluated. It is therefore relevant to take a step back from looking at the value of public transport for the individual and instead take a broader perspective and try to estimate the value from a society point of view.

The question of how land and housing prices relates to public transport is investigated to a wast extent in literature and articles. In the article *"Financing Transit Systems Through Value Capture"* (Smith and Gihring 2006) the authors summarize findings of over 100 studies concerning the impact of nearby public transport ("proximity to transit") on property values. The result they present indicates that "...proximity to transit often increases property values enough to offset some or all of transit system capital costs." On the contrary, another study in the Tyne and Wear region in the United Kingdom concludes, using a geographically weighted regression, that "...transport accessibility may have a positive effect on land value in some areas but a negative or no effect in others." (Hongbo Du 2006) In another study from The University of Hong Kong, the authors find that "accessibility to minibuses emerges as the most influential in determining house price". (H.M. So and Ganesan 1997) Yet another study evaluates the impact of the South Yorkshire Supertram on house prices in Sheffield with the result "Supertram initially depressed the prices of nearby houses but prices show signs of recovery." (Henneberry 1998)

From analysing earlier research on the subject of land and property values in relation to public transport we find that no perfect conclusions, satisfying all earlier studies, can be drawn. Instead, what stands out is rather the uniqueness of each solution, or as one may call it - *technical solution.* We recall here to Thomas P. Hughes theories about *Large Technical*

Systems which are both "socially constructed and society shaping" (Wiebe E. Bijker 1987) as well as Utterback and Abernathy's theories of *dominant design* to advocate for a theory that the development of public transport cannot simply be explained by common supply and demand economics but is rather heavily trajectorised by the technical systems and dominant designs surrounding it.

Technical systems differ from each other between cities and areas and thus pose different opportunities and solution possibilities. One example is Gärdet in Stockholm which at the moment is under reconstruction and according to the Swedish Traffic Board, Stockholms Läns Landsting is now running a project to combine the tramways Lidingö Tram (Swedish: Lidingöbanan) with Tram City (Swedish: Spårväg City), see figure A.1 in the Appendix. (Trafiknämnden 2011) This project has an estimated budget of 5-5.4 billion SEK and is scheduled to be finished in 2017. The decision is said to be based in a need for a more competitive public transport system providing the new district of Norra Djurgården in connection with Gärdet. We would like to argue that the decision likewise is based upon the possibilities that the present technical system provides in the area. Such aspects include both geographical constraints and the present solutions of public transport in the area and in neighboring areas. Another public transport solution which has been discussed for Stockholm is the building of a new metro line, "The Purple Line". This suggestion has been evaluated in a thesis at KTH Royal Institute of Technology in Stockholm with the conclusion that it is of utmost importance that traffic and city planning are not run as separate projects (Marcus Janson 2013). This conclusion supports our theory that the development of public transport is something that, in developed cities, is subject to technical systems and dominant designs. For these large systems the dominant design is not a subject to market competition but rather a subject to government regulations, or to quote Melissa A. Schilling in her book *Strategic Management of Technological Innovation* (Schilling 2013):

> "Where government regulation imposes a single standard on an industry, the technology design embodied in that standard necessarily dominates the other technology options available to the industry."

We can now recall our question: *Is proximity to public transport an important aspect when valuing apartments?* The answer we would like to propose is that on an individual level, you can have preferences for what attributes you seek when buying an apartment. However, if your preference is bound to a certain area that has good public transport; then you cannot, because of the tragedy of the commons, argue for price differentiating due to individual preferences. The development of public transport in an area is subject to large technical systems and thus it becomes irrelevant to try to put the value of it in a monetary context. Therefore, one should be careful when including a variable for proximity to public transport since it depends to a large extent on the application of the model.

Bibliography

Blume, Elin, Annica Streiler, and Henrik Weston (2013). *Läget i länet: Bostadsmarknaden i Stockholms län 2013*. Tech. rep. 14. Available at `http://www.lansstyrelsen.se/stockholm/SiteCollectionDocuments/Sv/publikationer/2013/rapport-2013-14.pdf`. Länsstyrelsen i Stockholms län.

Enheten för samhällsplanering (2014). *Bostadsmarknadsenkäten, Stockholms län*. Tech. rep. 8. Available at `http://www.lansstyrelsen.se/stockholm/SiteCollectionDocuments/Sv/publikationer/2014/rapport-2014-8.pdf`. Länsstyrelsen i Stockholms län.

Gunnvald, Rickard and Patrik Gunnvald (2012). "Estimation av bostadsrättspriser i Stockholms innerstad medelst multipel regressionsanalys (Swedish) [Estimation of apartment prices in the inner city of Stockholm using multiple regression analysis]". Bachelor thesis. Department of Mathematics, Royal Institute of Technology.

Hardin, Garrett (1968). "The Tragedy of the Commons". In: *Science* 162.

Hemnet (2014). URL: `http://www.hemnet.se/` (visited on 04/24/2014).

Henneberry, John (1998). "Transport investment and house prices". In: *Journal of Property Valuation and Investment* 16.2.

H.M. So, R.Y.C. Tse and S. Ganesan (1997). "Estimating the influence of transport on house prices: evidence from Hong Kong". In: *Journal of Property Valuation Investment* 15.1, pp. 40–47.

Hongbo Du, Corinne Mulley (2006). *Relationship Between Transport Accessibility and Land Value: Local Model Approach with Geographically Weighted Regression*. Tech. rep. Available at `http://trb.metapress.com/content/r4511852643v532m/`. Transportation Research Board.

Jennische, Andreas (2014). "Miljöpartiet går till val på 150 000 nya bostäder". In: *Direktpress - Östermalmsnytt*. URL: `http://www.direktpress.se/ostermalmsnytt/Nyheter1/Miljopartiet-gar-till-val-pa-150-000-bostader-till-2030/` (visited on 04/04/2014).

Kennedy, Peter (2008). *A Guide to Econometrics 6^{ed}*. Malden: Wiley-Blackwell.

Kollektivtrafik, Svensk (2012). *Investeringar för ett starkt och hållbart transportsystem - infrastrukturpropositionen prop. 2012/13:25*. Tech. rep. Svensk Kollektivtrafik.

Kuhn, Thomas S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lang, Harald (2013). *Topics on Applied Mathematical Statistics*. Stockholm: KTH.

Mäklarstatistik, Svensk (2014). URL: `http://www.maklarstatistik.se/` (visited on 04/24/2014).

Maps, Google (2014). URL: `https://www.google.se/maps/` (visited on 04/24/2014).

Marcus Janson, Tesad Alam (2013). "Lila Linjen för ett tätare Stockholm". Master Thesis. Institutionen för Samhällsplanering och Miljö, Royal Institute of Technology.

MathWorld, Wolfram (2014). URL: `http://mathworld.wolfram.com/HypothesisTesting.html` (visited on 04/28/2014).

Mayntz, Renate and Thomas P. Hughes (1988). *The Development of Large Technical Systems*. Boulder, Colorado: Westview Press.

Richard A. DeFusco et. al. (2007). *Quantitative Investment Analysis*. Hoboken, New Yersey: John Wiley & Sons.

Richard M. Heiberger, Burt Holland (2004). *Statistical Analysis and Data Display*. Philadelphia: Springer.

Salon, Deborah et al (2014). *The Impact of Bus Rapid Transit and Metro Rail on Property Values in Guangzhou, China*. Tech. rep. Available at `http://trid.trb.org/view.aspx?id=1290068`. Transportation Research Board.

SCB (2013). "Stockholmarna trivs trots bostadsbrist (Swedish) [Stockholm citizens thrive despite lack of housing]". In: *Välfärd (Swedish) [Welfare]* 4. Available at `http://www.scb.se/Statistik/_Publikationer/LE0001_2013K04_TI_00_A05TI1304.pdf`, pp. 8–9.

Schilling, Melissa A. (2013). *Strategic Management of Technical Innovation, 4^{th}ed.* New York: McGraw-Hill.

Sigot, Adrian (2014). Mailconversation. URL: `http://slutpris.se`.

Smith, Jeffery J. and Thomas A. Gihring (2006). "Financing Transit Systems Through Value Capture: An Annotated Bibliography". In: *American Journal of Economics an Sociology* 65.3.

Strand, J. and M. Vågnes (2001). "The relationship between property values and railroad proximity: a study based on hedonic prices and real estate brokers' appraisals". In: *Transportation* 28.2, pp. 137–156. DOI: `10.1023/A:1010396902050`.

Svensk Mäklarstatistik AB (2014). *Statistik Stockholm*. URL: `http://www.maklarstatistik.se/maeklarstatistik/laen.aspx?LK=180&Typ=Boratter&srt=asc&tab=namn` (visited on 03/30/2014).

Trafiknämnden (2011). *Genomförandebeslut - Spårväg City, fortsatt utbyggnad*. Tech. rep. Stockholms Läns Landsting.

Tukey, John W. (1986). *The Collected Works of John W. Tukey - Philosphy and Principles of Data Analysis*. Monterey, California: Wadsworth Brooks/ Cole.

Valueguard (2014). *Nasdaq OMX Valueguard-KTH Housing Index (HOX) Stockholm BR*. URL: `http://www.valueguard.se/stockholmbr` (visited on 04/20/2014).

Wiebe E. Bijker Thomas Parke Hughes, Trevor J. Pinch (1987). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT: MIT Press.

A Appendices

Table A.1: Coefficient covariance matrix estimate from OLS before using White's robust estimate. All values times 10^9.

2.34	0.00	-0.12	-1.16	-1.09	-0.89	-0.67	-0.70	-0.61	-0.65	-0.66	-0.26	0.10	-0.05	0.00	0.09
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-0.12	0.00	0.21	0.05	0.05	0.01	0.02	-0.02	0.00	-0.04	0.00	-0.01	-0.02	0.03	0.00	-0.02
-1.16	0.00	0.05	1.16	0.90	0.84	0.07	0.12	-0.01	0.11	0.08	0.07	-0.15	-0.02	0.00	-0.06
-1.09	0.00	0.05	0.90	0.97	0.82	0.06	0.04	0.00	-0.01	0.03	0.03	-0.09	0.00	0.00	-0.05
-0.89	0.00	0.01	0.84	0.82	1.10	0.06	0.06	0.00	-0.01	-0.01	0.00	-0.01	0.00	0.00	0.00
-0.67	0.00	0.02	0.07	0.06	0.06	2.17	0.60	0.58	0.59	0.60	0.01	0.02	0.04	0.00	-0.04
-0.70	0.00	-0.02	0.12	0.04	0.06	0.60	0.81	0.58	0.64	0.62	0.01	0.05	0.04	0.00	-0.02
-0.61	0.00	0.00	-0.01	0.00	0.00	0.58	0.58	0.80	0.57	0.58	0.00	0.04	0.03	0.00	-0.03
-0.65	0.00	-0.04	0.11	-0.01	-0.01	0.59	0.64	0.57	1.09	0.63	0.02	0.00	0.00	0.00	0.01
-0.66	0.00	0.00	0.08	0.03	-0.01	0.60	0.62	0.58	0.63	0.75	0.00	0.05	0.02	0.00	-0.01
-0.26	0.00	-0.01	0.07	0.03	0.00	0.01	0.01	0.00	0.02	0.00	0.33	0.00	-0.08	0.00	0.00
0.10	0.00	-0.02	-0.15	-0.09	-0.01	0.02	0.05	0.04	0.00	0.05	0.00	0.46	0.00	0.00	-0.05
-0.05	0.00	0.03	-0.02	0.00	0.00	0.04	0.04	0.03	0.00	0.02	-0.08	0.00	0.30	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.09	0.00	-0.02	-0.06	-0.05	0.00	-0.04	-0.02	-0.03	0.01	-0.01	0.00	-0.05	0.00	0.00	1.50

Table A.2: Coefficient covariance matrix estimate from OLS using White's robust estimate. All values times 10^9.

3.49	-0.01	0.00	-1.83	-1.69	-1.35	-1.05	-1.04	-0.89	-1.02	-0.94	-0.24	0.25	-0.12	0.00	0.51
-0.01	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.01	0.00	-0.01	0.00	-0.01	0.00	0.00	0.00
0.00	0.00	0.18	0.00	0.02	0.03	-0.04	-0.02	0.01	-0.06	0.01	0.00	0.04	0.01	0.00	-0.01
-1.83	0.00	0.00	1.74	1.45	1.33	-0.05	0.07	-0.12	0.02	-0.01	0.07	-0.25	0.03	0.00	-0.24
-1.69	0.00	0.02	1.45	1.51	1.31	-0.09	-0.12	-0.21	-0.23	-0.18	0.01	-0.11	0.06	0.00	-0.22
-1.35	-0.01	0.03	1.33	1.31	1.72	-0.02	0.01	-0.05	-0.15	-0.08	-0.01	-0.01	0.07	0.00	-0.07
-1.05	-0.01	-0.04	-0.05	-0.09	-0.02	3.76	1.51	1.55	1.52	1.53	0.02	-0.01	0.02	0.00	-0.16
-1.04	-0.01	-0.02	0.07	-0.12	0.01	1.51	1.59	1.47	1.51	1.47	0.05	0.01	0.05	0.00	-0.07
-0.89	-0.01	0.01	-0.12	-0.21	-0.05	1.55	1.47	1.72	1.48	1.49	0.05	0.08	0.04	0.00	-0.08
-1.02	0.00	-0.06	0.02	-0.23	-0.15	1.52	1.51	1.48	2.06	1.51	0.06	-0.05	-0.02	0.00	-0.03
-0.94	-0.01	0.01	-0.01	-0.18	-0.08	1.53	1.47	1.49	1.51	1.60	0.04	0.03	0.04	0.00	-0.06
-0.24	0.00	0.00	0.07	0.01	-0.01	0.02	0.05	0.05	0.06	0.04	0.29	0.01	-0.09	0.00	-0.15
0.25	-0.01	0.04	-0.25	-0.11	-0.01	-0.01	0.01	0.08	-0.05	0.03	0.01	0.74	-0.02	0.00	-0.12
-0.12	0.00	0.01	0.03	0.06	0.07	0.02	0.05	0.04	-0.02	0.04	-0.09	-0.02	0.30	0.00	-0.03
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.51	0.00	-0.01	-0.24	-0.22	-0.07	-0.16	-0.07	-0.08	-0.03	-0.06	-0.15	-0.12	-0.03	0.00	3.69

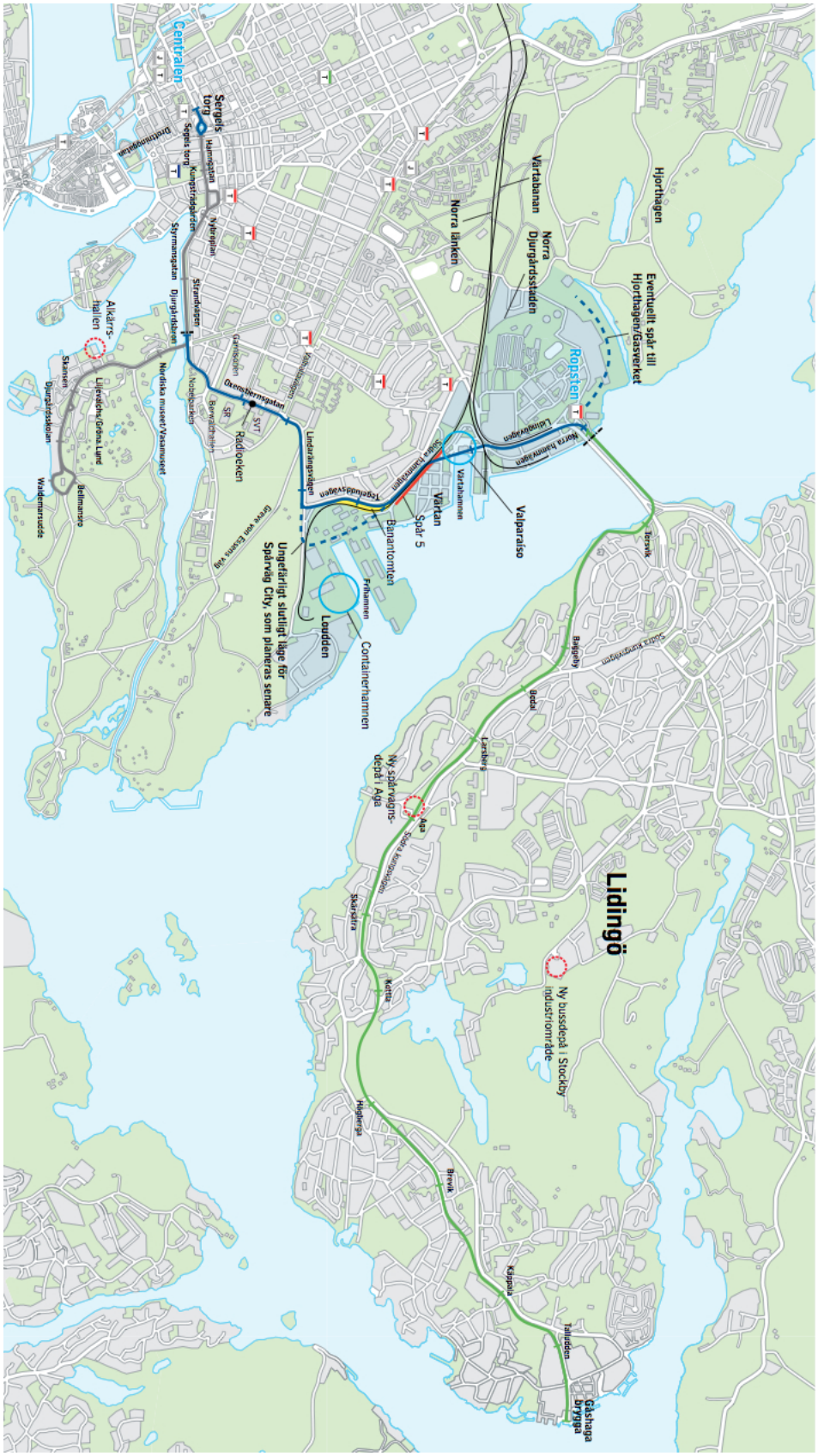


Figure A.1: Map of planed extensions of the City Tram.

TRITA -MAT-K 2014:06
ISRN -KTH/MAT/K--14/06-SE