Statistical learning procedures for analysis of residential property price indexes

Rydén, Otto
oryden@kth.se

SF290X Degree Project in Mathematical Statistics, Second Cycle
Department of Mathematical Statistics
KTH Royal Institute of Technology, Stockholm

Supervisor:

Almér, Henrik
henrik.almer@booli.se

Pavlenko, Tatjana
pavlenko@math.kth.se

Examiner:

Pavlenko, Tatjana
pavlenko@math.kth.se

May 16, 2017

This page is intentionally left blank

Abstract

Residential Price Property Indexes (RPPIs) are used to study the price development of residential property over time. Modeling and analysing an RPPI is not straightforward due to residential property being a heterogeneous good. This thesis focuses on analysing the properties of the two most conventional hedonic index modeling approaches, the hedonic time dummy method and the hedonic imputation method. These two methods are analysed with statistical learning procedures from a regression perspective, specifically, ordinary least squares regression, and a number of more advanced regression approaches, Huber regression, lasso regression, ridge regression and principal component regression. The analysis is based on the data from 56 000 apartment transactions in Stockholm during the period 2013-2016 and results in several models of a RPPI. These suggested models are then validated using both qualitative and quantitative methods, specifically a bootstrap re-sampling to perform analyses of an empirical confidence interval for the index values and a mean squared errors analysis of the different index periods. Main results of this thesis show that the hedonic time dummy index methodology produces indexes with smaller variances and more robust indexes for smaller datasets. It is further shown that modeling of RPPIs with robust regression generally results in a more stable index that is less affected by outliers in the underlying transaction data. This type of robust regression strategy is therefore recommended for a commercial implementation of an RPPI.

Sammanfattning

Bostadsprisindex används för att undersöka prisutvecklingen för bostäder över tid. Att modellera ett bostadsprisindex är inte alltid lätt då bostäder är en heterogen vara. Denna uppsats analyserar skillnaden mellan de två huvudsakliga hedoniska indexmodelleringsmetoderna, som är, hedoniska tid-dummyvariabelmetoden och den hedoniska imputeringsmetoden. Dessa metoder analyseras med en statistisk inlärningsprocedur gjord utifrån ett regressionsperspektiv, som inkluderar analys utav minsta kvadrats-regression, Huberregression, lassoregression, ridgeregression och principal component-regression. Denna analys är baserad på ca 56 000 lägenhetstransaktioner för lägenheter i Stockholm under perioden 2013-2016 och används för att modellera flera versioner av ett bostadsprisindex. De modellerade bostadsprisindexen analyseras sedan med hjälp utav både kvalitativa och kvantitativa metoder inklusive en version av bootstrap för att räkna ut ett empiriskt konfidensintervall för bostadsprisindexen samt en medelfelsanalys av indexpunktskattningarna i varje tidsperiod. Denna analys visar att den hedoniska tid-dummyvariabelmetoden producerar bostadsprisindex med mindre varians och ger också robustare bostadsprisindex för en mindre datamängd. Denna uppsats visar också att användandet av robustare regressionsmetoder leder till stabilare bostadsprisindex som är mindre påverkade av extremvärden, därför rekommenderas robusta regressionsmetoder för en kommersiell implementering av ett bostadsprisindex.

Acknowledgements

I would like to thank my thesis supervisor, Tatjana Pavlenko, Associate Professor at the Department of Mathematics at KTH Royal Institute of Technology for her encouragement throughout the intense work of finishing this thesis. I would also like to thank Booli Search Technologies AB and especially my supervisor Henrik Almér for providing me with the data and the interesting problem investigated in this thesis. I also want to thank my peers at KTH for five fantastic and intensive years and for all the help that they have provided during the creation of this thesis.

Stockholm, May 2017
Otto Rydén

Contents

List of Figures

List of Tables

Abbreviations

SS	***S***um of ***S***quares
PC	***P***rincipal ***C***omponent
PCR	***P***rincipal ***C***omponent ***R***egression
OLS	***O***rdinary ***L***east ***S***quares
BLUES	***B***est ***L***inear ***U***nbiased ***ES***timator
AIC	***A***kaike ***I***nformation ***C***riterion
BIC	***B***ayesian ***I***nformation ***C***riterion
LAD	***L***east ***A***bsolute ***D***eviation
RPPI	***R***esidential ***P***roperty ***P***rice ***I***ndex
HDI	***H***edonic ***D***ouble ***I***mputation
MSE	***M***ean ***S***quared ***E***rror

Symbols/Notations

$\boldsymbol{y}$	Regression vector
$\hat{\boldsymbol{y}}$	Predicted regression vector
$\bar{x}$	Mean value of vector x
$\boldsymbol{X}$	A matrix containing all independent variables
$\boldsymbol{X_i}$	The i:th column in $\boldsymbol{X}$
$\boldsymbol{\beta}, \boldsymbol{\gamma}$	Coefficient vectors for regression
β_i, γ_i	The i:th coefficient in $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$
$\hat{\boldsymbol{\beta}}_M, \hat{\boldsymbol{\gamma}}_M$	The estimated coefficient vector using method M
$\hat{\beta}_i, \hat{\gamma}_i$	The i:th coefficient in $\hat{\boldsymbol{\beta}}_M$, $\hat{\boldsymbol{\gamma}}_M$
ϵ_i	The error term from a regression
e_i	The residual from a regression ($e_i = y_i - \hat{y_i}$)
σ	The standard deviation
f	Degrees of freedom
p	The numbers of covariates
n	The numbers of data points
s	The estimated standard error, $s = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$
$SS_{res}(p)$	The residual sum of squares for a regresion model with $p-1$ regressors and an intercept. $SS_{res}(p) = \boldsymbol{y}^\intercal\boldsymbol{y} - \hat{\boldsymbol{\beta}}^\intercal\boldsymbol{X}^\intercal\boldsymbol{y}$
SS_T	The total sum of squares. $SS_T = \boldsymbol{y}^\intercal\boldsymbol{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$
$E(X)$	The expected value for a random variable X
$Var(X)$	The variance for a random variable X
$cor(X, Y)$	the correlation between the random variables X and Y
$\log(X)$	The natural logarithm of X
$\lvert\boldsymbol{a}\rvert$	The norm of vector $\boldsymbol{a}$
$\boldsymbol{X}^\intercal$	The transpose of matrix $\boldsymbol{X}$
$\boldsymbol{I}$	The identity matrix
p_n^t	price of property n at time t
P_M^{0t}	A index defined from time 0 to time t using the index method M

1 Introduction

In this section we are first going to provide a brief background to the main underlying problems. We then introduce the purpose and research question before the delimitations and limitations are presented. The section ends with a short passage about the expected contribution of this thesis and an outline for the rest of the thesis is then presented.

1.1 Background and problem formulation

The general price trend in the residential property market has implications on many areas of today's society including the cost of living and ones willingness to move. In this thesis we will therefore study methods to model the price levels in the housing market. The price level will be assessed by modeling a residential property price index (RPPI) for Stockholm during the period 2013-2016 using statistical learning procedures from a regression perspective. The problem originates from the intent to get a fair value of a property sold in the housing market. In areas with a high transaction volume (usually cities) similar objects will be sold frequently and therefore similar newly sold objects can be used for valuing the object for sale but in areas with a lower transaction volume fewer recently sold objects can be used to value an object for sale. The price level index would then be used to move all objects sold in the past to the same point in time so that they could be used for valuation of an object.

The difficulties with modeling a RPPI comes from the fact that residential property is a heterogeneous good. Modelling of a price index for a homogeneous good (like gold or oil) is easier because many transactions of identical goods are performed frequently, however for a heterogeneous good it is harder to describe a change in the price level if two different objects are sold in two different time periods. One could use an average of the objects sold during two different time periods to create an index but it would lead to a volatile and inexact index. The modeling of indexes for heterogeneous goods therefore consists of ways to remove the effect from these differences in the underlying characteristics.

1.2 Purpose

There exists much literature on how to model RPPIs and this thesis will use statistical learning procedures to model and analyse the different methodologies described in this literature. Valueguard already produces RPPIs (The HOX-index) for the Swedish market but this thesis will compare the hedonic time dummy method that is used in the modeling of their indexes with other methods (mainly the hedonic double imputation method) and see under which circumstances the different methods perform better. The thesis will focus on the statistical properties of the index modeling and on methods to test the proficiency of the different index models.

1.3 Research questions

We have formulated the following research questions:

RQ1: Which/What index methodology works best for modeling an index that captures the price trend for apartments?

RQ2: What different ways are there to analyse the robustness of the constructed indexes?

More specifically, the focus will be on the following goals:

- Investigate which statistical methods are the best ones for modeling an index for heterogeneous objects
- Model a variety of RPPI's using statistical learning perspectives from an regression approach and further evaluate the indexes using computing intensive statistical methods.
- Improve the accuracy of Booli's pricing algorithm by usage of a more accurate modeling of RPPIs.

1.4 Delimitation

The thesis will only investigate residential properties and not other forms of property like industrial property or office houses. This thesis is only focused on the Stockholm market even if the main methodology should be applicable to the whole of Sweden and to similar countries.

The focus of this thesis is on a hedonic regression model approach and will mainly focus on the differences between the hedonic time dummy approach and the double imputation approach. Other methods included stratification, repeat sales and assessment based pricing methods will not be investigated.

This thesis does not investigate non-linear models for index creation as initial analysis of non-linear model pointed at unsatisfactory results. Most of the index creation models discussed in the literature are linear models and an introduction of non-linear models would only create a more complex model which is harder to implement and interpret.

1.5 Limitation

This study only handles data for the period of 2013-2016 as the used dataset does not contain any older transaction data. Another limitation of this study is that the dataset used did not contain any information that could be used to validate the exactness of the calculated RPPIs. The data source used in this report does not contain some characteristics that could be beneficial in a valuation model and therefore in modeling of a RPPI.

1.6 Outline

The outline of the rest of the report will now be visualized with an aim to give the reader a better understanding of the outline of the report.

Section 2: This section presents the theoretical background needed to understand the rest of the report. The chapter starts with describing the standard linear model and then describes more advanced regression methods used for fitting the linear model including; OLS regression, Ridge regression, Lasso Regression, PCR and M-estimation (Huber regression).

Section 3: This section presents a literature review and present the current state of knowledge in index modeling methods. The section start by providing a short summary of the papers and books that handles index modeling before the current Swedish RPPI (the HOX-index) is described.

Section 4: In this section we present the data used in the modeling and analysis of the RPPIs. The section starts by describing how the data were collected and which characteristics each data point contains. The monthly data generating process and the amount of missing data is then addressed.

Section 5: This section presents the methodology used in the modeling and validation of the RPPIs. The section begins with a description of the different index methodologies before the implementation and validation procedure is described.

Section 6: This section handles the results and analysis performed in this thesis. It starts by analysing a valuation modell which is going to be used in the index modelling. RPPIs is then modeled using both the hedonic time dummy method and the hedonic double imputation method. The section ends with the application of different validation methods on the modeled RPPIs.

Section 7: This section ends the report with a summary of the previous results and analyses. Furthermore, research topics are then suggested before a recommendation for a commercial implementation of a RPPI is provided.

2 Theory: Statistical learning from the perspective of linear regression models

We shall here present some of the index methods and the statistical theory used in the paper. This section is an overview and an interested reader should check the references for a more thorough description of the theory. Most of the theory is about linear regression models but some theory about index modeling is also included.

2.1 Linear model

Most of the methodology in this thesis will be based on the linear model so we start with a definition of the basic linear model:

$$p_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \epsilon_n^t \tag{2.1}$$

where we use the following notions:
p_n^t denotes the price of property n at time t
z_{nk}^t denotes the value of "quality" k for property n at time t
β_0^t and β_k^t denotes the intercept term and the characteristic parameters to be estimated
ϵ_n^t denotes the error term of property n at time t

The linear model in equation (2.1) describes how the price p_n^t can be described by a linear combination of some of the K characteristics of the property plus the error term ϵ_n^t, which is the difference between the real value p_n^t and the predicted value $\hat{p}_n^t$.

We are now going to present a useful property of the linear model which states that the valuation in a linear valuation model becomes the same if one takes the average of the independent variables as an input and the average of the dependent variables with all the individual independent variables as input. We have the following linear model for one object n of the dependent variable y_n:

$$y_n = \beta_0 + \sum_{k=1}^{K} \beta_k z_{nk} + \epsilon_n \tag{2.2}$$

one therefore gets the mean value $\bar{y}$ for N dependent variables y_n by the following formula that also shows the statement made above.

$$\bar{y} = \frac{\sum_{n=1}^{N}(\beta_0 + \sum_{k=1}^{K}\beta_k z_{nk} + \epsilon_n)}{N} = \beta_0 + \frac{\sum_{k=1}^{K}\beta_k \sum_{n=1}^{N}(z_{nk}) + \sum_{n=1}^{N}(\epsilon_n)}{N} =$$
$$\beta_0 + \sum_{k=1}^{K}\beta_k \frac{\sum_{n=1}^{N}(z_{nk})}{N} + \frac{\sum_{n=1}^{N}(\epsilon_n)}{N} = \beta_0 + \sum_{k=1}^{K}\beta_k(\bar{z}_k) + \frac{\sum_{n=1}^{N}(\epsilon_n)}{N} \tag{2.3}$$

For imputation one does not use the error term so the last term in equation (2.3) will disappear. Hence, we see that for imputation it does not matter if one takes the average of the dependent variables or the independent ones.

2.2 Regression models

For a given linear regression model we use various estimations to obtain the parameters and we will now present some different parameter estimation methods. We start with the standard Ordinary Least Squares (OLS) parameter estimation method and describe some properties and problems of this method. We then introduce parameter estimation methods that are used to better handle some of the problems that can occur with the OLS method. These methods include; Ridge regression, Lasso regression, PCR (principal component regression) and Huber regression.

2.2.1 OLS regression

The OLS regression estimation approach minimizes the residual sum of squares. We use the following notation for the training data [17]:

$$(\mathbf{X}^i, y_i) \quad i = 1, 2, ..., N \qquad \mathbf{X}^i = (x_{i1}, ..., x_{ip}) \tag{2.4}$$

where y_i represents the generic dependent variable used instead of p_n^t in equation (2.1). We get $\hat{\boldsymbol{\beta}}_{OLS} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ from the following equation

$$\hat{\boldsymbol{\beta}}_{OLS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j}^{p} \beta_j x_{ij})^2\} \tag{2.5}$$

using matrix notations we get the following solution of equation (2.5): [10]

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X} \mathbf{y} \tag{2.6}$$

where $\mathbf{X} = \begin{pmatrix} 1 & \mathbf{X}^1 \\ \vdots & \vdots \\ 1 & \mathbf{X}^N \end{pmatrix}$ is the $N \times (p+1)$ matrix containing all the independent variable and the intercept term 1 and $\mathbf{y} = (y_1, ..., y_N)^\intercal$ is the $N \times 1$ vector containing the dependent variables.

We have made five assumptions for the regression analysis to hold. If we break these assumptions we are not able to perform the statistical tests of the coefficient intervals. The assumptions are: [10]

1. The relationship between the response variable y and the independent variables x_i are at least approximately linear.
2. $E[\epsilon_i] = 0$ which means that the error term of the regression has zero mean.
3. The variance of the error term ϵ_i is constant for different values of the dependent variable y.
4. $cor(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$ which means that the error terms are uncorrelated.
5. The error term has a normal distribution.

If the error term ϵ_i in the OLS regression does not have a constant variance for different values of y the OLS regression will still be unbiased but the model will not have the minimum variance property of BLUES (Best Linear Unbiased EStimator).

Under the Gauss-Markov assumptions (assumption 1-4) the OLS regression is BLUES , it is therefore the unbiased estimator with the smallest variance of the coefficients. However to require that the estimator is unbiased is a large restriction and if one allows some bias one can find estimators with smaller variance [10].

2.2.2 Methods for model selection (for OLS)

A very important part of building a linear regression model is the choice of regressors to include in the model. This choice will affect the usefulness of the model so we are now going to present different measures to evaluate a specific selection of variables for a linear model. There is no perfect measure for deciding which subset of variables on should use, but there are some statistics which are frequently used for variable determination. We start by describing the coefficient of multiple determination also known as the R^2-value. The R^2 value for a model with $p-1$ regressor terms and an intercept term is calculated using the following formula:

$$R_p^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{Res}(p)}{SS_T} \tag{2.7}$$

One problem with the R^2-value is that it is an increasing function with p so more regressors give a higher R^2-value. One can therefore use the adjusted R^2-value that takes this problem into account.

$$R^2_{Adj,p} = 1 - (\frac{n-1}{n-p})(1 - R^2_p) \tag{2.8}$$

A criterion that one could use for model selection is the model that maximize $R^2_{Adj,p}$[10].

Another measure for choosing a model is the residual mean square:

$$MS_{Res}(p) = \frac{SS_{res}(p)}{n-p} \tag{2.9}$$

one can show that the model that minimizes the $MS_{Res}(p)$ also maximizes $R^2_{Adj,p}$[10].

Two other methods for model selection is Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), one then chooses the model which has the lowest AIC or BIC value. The AIC is based on a maximization of the entropy of the model, it also is a log-likelihood measure with a penalizing term for many variables:

$$AIC = -2\log(L) + 2p \tag{2.10}$$

in the OLS model the log likelihood function takes the form $L = \frac{SS_{Res}}{n}$ which gives us:

$$AIC = -2\log(\frac{SS_{Res}}{n}) + 2p \tag{2.11}$$

The BIC number builds on the same idea but adds more penalty for adding more regressors as the sample size grows.

$$BIC = -2\log(L) + p\log(n) \tag{2.12}$$

In the OLS model the log likelihood function takes the form $L = \frac{SS_{Res}}{n}$ which gives us:

$$BIC = -2\log(\frac{SS_{Res}}{n}) + p\log(n) \tag{2.13}$$

2.2.3 Variable transformations the OLS

As described above, the estimated regression demonstrates poor performance if the choice of regressors in the model is not right. But even if the model contains the right regressors there can still be problems if the dependent variable does not have a linear relationship with the regressors. In some cases the problem with a non-linear function can be solved by linearisation using a transformation, those linear models are called *transformable* or *intrinsically* linear models. One example of an transformable linear model is a model where the relationship between the dependent and independent variables are exponential, then taking the natural logarithm of the dependant variable would lead to a linearised model [10]. This problem can be solved by transforming the dependent variable which will be described next.

If the data does not have a normal distribution and if the error terms ϵ_i have different values for the variance one could transform the dependent variable in the model to get a better model. One type of transformation is the class of power transformations y^λ, where we want to decide the best value for the transformation parameter λ and this is done with the maximum likelihood method. If one use the transformation y^λ there will be problems when $\lambda = 0$ and this is solved by using the following transformation named Box-Cox transformation (see equation (2.14)):

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y}\log(y) & \lambda = 0 \end{cases} \tag{2.14}$$

where $\dot{y} = \log^{-1}[l/n \sum_{i=1}^{n} \log(y_i)]$ is the geometric mean of the observations.

The maximum likelihood estimation of λ in equation (2.14) corresponds to the value of λ that leads to a minimum value of the fitted model's sum of squared residuals [5]. One therefore usually splits the range of λ into a grid and calculates the maximum likelihood value for the different λ values. This is done with the boxcox function in R. One usually calculates an approximate confidence interval for λ (See [10] page 183) to justify choosing a nicer value for λ (choosing $\lambda = 1$ instead of $\lambda = 0.95$ if 1 is in the confidence interval for λ) [10].

If we transform the dependent variable using the natural logarithm and use this equation to fit the model

$$\log(p_n^t) = \beta_0^t + \sum_{k=1}^{p} \beta_k^t z_{nk}^t + \epsilon_n^t \tag{2.15}$$

and then use the inverse to the log function (the exponential function) to estimate the value of the dependent variable p_n^t the estimation would be biased. This bias can be corrected by using the following formula

$$p_n^t = exp(\log(p_n^{t*}) + s^2/2) \tag{2.16}$$

where we have that s^2 is the unbiased estimator of σ^2 (the variance of the residual terms ϵ) in equation (2.15). But equation (2.16) is only an unbiased estimator of p_n^t if the errors ϵ_n^t in equation (2.15) are normally distributed [6].

2.2.4 Influential point analysis for the OLS model

A data set that contains outliers can create problems for a regression model. Outlier points can effect the parameters in the model and lead to an unrobust model. We will present three different ways (Cook's distance, DFFITS and DFBETAS) to detect influential points in this subsection.

To decide how much influence one data point has on the regression model one usually calculates the Hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$. The element h_{ij} in the hat matrix $\mathbf{H}$ can then be interpreted as the amount of leverage that observation y_i has on the fitted value $\hat{y}_i$. One gets that the average value $\bar{h} = p/n$ and a point that is twice the average value is considered to be a leverage point (a point with a value over $2p/n$) [10].

Not all leverage points need to be influential points and there exist some tests to see if a point is considered to be an influential point, one of these tests is Cook's D which has the general form [10]:

$$D_i(\mathbf{M}, c) = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^\intercal \mathbf{M} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{c} \quad i = 1, 2, ..., n \tag{2.17}$$

One usually sets $\mathbf{M} = \mathbf{X}^\intercal\mathbf{X}$ and $c = p * MS_{Res}$ which give the following formula for Cook's D

$$D_i(\mathbf{M}, c) = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^\intercal \mathbf{X}^\intercal\mathbf{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p * MS_{Res}} \quad i = 1, 2, ..., n \tag{2.18}$$

There exist many different ways to interpret if a point is an influential point according to the Cook's D measure. One way is to compare D_i with the α-quartile of the F-distribution $F_{\alpha,p,n-p}$. If the calculated $D_i = F_{0.5,p,n-p}$ deleting the point would correspond to moving $\hat{\boldsymbol{\beta}}_{(i)}$ to the boundary of a 50% confidence region for $\boldsymbol{\beta}$ which is based on the whole dataset. This indicates that the OLS estimate is sensitive to data point i. We have that $F_{0.5,p,n-p} \approx 1$ thus according to this measure one considers $D_i > 1$ to be a sign of an influential point [10]. The previous described cut-off value of Cook's D is high if the data sample is large and another cut-off measure for Cook's D is when $D_i > \frac{4}{n}$ which gives it a similar behaviour as for DFFITS cut-off value described below [3].

DFBETAS indicates how many standard deviation the regression coefficient $\hat{\beta}_j$ changes if the i:th observation were deleted.

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}} \tag{2.19}$$

where $S_{(i)}$ is the standard error estimated without the point i in question, C_{jj} is the j:th diagonal element of $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ and $\hat{\beta}_{j(i)}$ is the j:th regressor coefficient calculated without the i:th observation.

DFFITS measures how many standard deviations the fitted value will change if the i:th point is deleted.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)}h_{ii}}} \tag{2.20}$$

where $S_{(i)}$ is the standard error estimated without the point i in question, $\hat{y}_{(i)}$ is the fitted value of $\hat{y}_i$ calculated without the i:th observation and h_{ii} is the i:th diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$.

The suggested cut-off value for these influential point measures is that one should examine a data point further if $|DFBETAS_{j,i}| > 2/\sqrt{n}$ or if $|DFFITS_i| > 2\sqrt{p/n}$ [10].

2.2.5 Multicollinearity

When one builds a regression model one hopes that the included regressors in that model are orthogonal, however when the linear dependence between the regressors are large the model suffers from a multicollinearity problem. There are many sources of multicollinearity but the two sources that are most relevant to this paper are multicollinearity originating from the model specification or multicollinearity originating from an over specified model.

Multicollinearity leads to large variances and covariances for the estimated regressor coefficient which leads to an unstable model. One way to detect multicollinearity is to calculate the variance inflation factors of the model:

$$VIF_j = \frac{1}{1 - R^2_j} \tag{2.21}$$

where R^2_j is the R^2 value of the regression with variable j as the dependent variable and the other independent variables as regressors (see equation (2.22))

$$x_i = \beta_0 + \sum_{k \neq i}^{K} \beta_k x_k + \epsilon_k \tag{2.22}$$

One usually says that the model suffers from high multicollinearity if some regressors have VIF values that exceeds 5 or 10. There are many different methods for dealing with a model with high multicollinearity, including collecting more data which would lead to lower variances of the coefficients. One could also respecify the model removing some of the independent variables that suffer from high multicollinearity and one could also use other methods than OLS. Another model that is used for handling multicollinearity is the ridge regression method which will be described next. [10].

2.3 Shrinkage regression methods

When the data follows the five assumptions mentioned earlier OLS is a good and reasonable choice of estimation method. However when some of the assumptions are violated or if the data contains outliers or is subject to high multicollinearity one should consider using more advanced regression methods. We will here describe shrinkage regression methods including ridge regression and lasso regression. Both the ridge regression and lasso regression methods are used to improve a model that suffers from high multicollinearity. However this section will start with a description of biased and unbiased estimators as shrinkage methods often leads to a biased model.

2.3.1 Biased/Unbiased estimators

We stated in the Gauss-Markov theorem that an OLS regression is BLUES which implies that it is the best linear unbiased estimator of the regression parameters $\hat{\boldsymbol{\beta}}$. Here this implies the estimator with the smallest variance of $\hat{\boldsymbol{\beta}}$. If we break down the MSE (Mean Squared Error) of an estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$.

$$MSE(\hat{\boldsymbol{\beta}}) = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = Var(\hat{\boldsymbol{\beta}}) + [E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}]^2 \tag{2.23}$$

So the mean square error of an estimator is the variance of the estimator plus the bias of the estimator, so in some instances one could choose to use a biased model in order to get a smaller variance and therefore smaller MSE [10].

2.3.2 Ridge regression

The ridge regression method minimizes the residual sum of squares subject to the restriction that the sum of the squares of the coefficients is less than a constant. [17]

$$\hat{\boldsymbol{\beta}}_R = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\} \quad \text{subject to } \sum_{j=1}^{p}\beta_j^2 \leq t \tag{2.24}$$

The data is standardized in the ridge regression so the magnitude of the variables does not affect the constraint of the model. The standardization of the variables also help comparisons of the coefficients in for example trace plots. The ridge regression model is a potential solution to a model that suffers from problems with multicollinearity as it put constrains on the included coefficients and hinder the estimated coefficients from becoming large and unrobust due to multicollinearity [10].

Another way to express equation (2.24) is to write it in a closed form with a penalty term λ:

$$\hat{\boldsymbol{\beta}}_R = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2\} \tag{2.25}$$

where the parameter λ in equation (2.25) and t in equation (2.24) are related to each other. Equation (2.25) has the solution [17]:

$$\hat{\boldsymbol{\beta}}_R = (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{y} \tag{2.26}$$

One can clearly see that equation (2.25) becomes an OLS regression equation when $\lambda = 0$. We have that the ridge regression parameter is a linear transformation of the least squares estimator and we will now check the bias of the ridge estimator by breaking down the mean square error to its components [10]:

$$\begin{aligned}\text{MSE}(\hat{\boldsymbol{\beta}}_R) = \text{Var}(\hat{\boldsymbol{\beta}}_R) + (\text{bias in } \hat{\boldsymbol{\beta}}_R)^2 = \\ \sigma^2\text{Trace}[(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}] + \lambda^2\hat{\boldsymbol{\beta}}^\intercal(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-2}\hat{\boldsymbol{\beta}} = \\ \sigma^2\sum_{j=1}^{p}\frac{h_j}{(h_j + \lambda)^2} + \lambda^2\hat{\boldsymbol{\beta}}^\intercal(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-2}\hat{\boldsymbol{\beta}}\end{aligned} \tag{2.27}$$

where $\hat{\boldsymbol{\beta}}$ is the parameter from the OLS regression equation on the same dataset and h_j is the eigenvalues of $\mathbf{X}^\intercal\mathbf{X}$. The first term on the right hand side in equation (2.27) can be seen as the sum of the variance of the parameters in $\hat{\boldsymbol{\beta}}_R$ and the second term can be seen as the squared bias. One therefore observes that the bias increases with increasing λ and that the variance decreases with increasing λ. This is called the Bias-variance trade-off strategy and it is important to choose a good value of λ which is done by cross validation in this thesis [10].

One usually finds the optimal model for different values of the restriction term λ and compare the models using cross validation to find the best restriction according to the mean squared error in the cross validation. The Glmnet package for R does this for 100 different values of λ [13].

2.3.3 Lasso regression

The lasso regression (least absolute shrinkage and selection operator) method minimizes the residual sum of squares subject to the restriction that the sum of the absolute value of the coefficients is less than a constant [17].

$$\hat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\} \quad \text{subject to } \sum_{j=1}^{p}|\beta_j| \leq t \tag{2.28}$$

As for ridge regression described above we can express equation (2.28) by writing it in a closed form with a penalty term λ [17]:

$$\hat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|\} \tag{2.29}$$

One can not write a solution in closed form for the lasso equation (2.29) and the solution is obtained by solving the quadratic programming problem that is stated in equation (2.28) [17].

One can note several similarities between the lasso regression method and the ridge regression method. The thing that differentiates the methods is that the Ridge regression method has a quadratic "penalty term" (see equation (2.25)) and that the lasso regression method has an absolute value "penalty term" (see equation (2.29)). These differences lead to some differences in the solutions of the models. In the lasso model some regressors are usually set to 0 while in the Ridge model these variables are usually very small but larger than 0 [17].

The same methodology with cross validation to choose λ that was described for the Ridge regression above is used to find the "best" lasso regression model. The R package Glmnet is also used for the lasso regression.

2.4 Regression methods using derived inputs

When a model have a large number of inputs, that often are very correlated, it could be beneficial to perform a regression of a linear combination of the independent variables instead on all of the independent variables. This section will describe the PCR (Principal Component Regression) which uses the principal components of the independent variables as regressors in the regression model.

2.4.1 PCR (Principal Component Regression)

The idea behind principal component regression or PCR is to calculate the principal components PC for the independent variables $\mathbf{X}$ and use these PCs to perform the regression on the dependent variable $\boldsymbol{y}$. One adds one PC at a time starting with the PCs with the largest explanation power of the variance [15].

There are many potential advantages with using a PCR instead of OLS regression including; dimensionality reduction, avoidance of multicollinearity between predictors and over fitting mitigation. One drawback with PCR is that one should not use PCR as a variable selection method as the usage of PCR can lead to difficulties of explaining which factor that affect the dependent variable [15].

The parameter vector $\hat{\boldsymbol{\beta}}_{PC}$ can be expressed by the following expression: [10]

$$\hat{\boldsymbol{\beta}}_{PC} = \boldsymbol{T}\hat{\boldsymbol{\alpha}}_{PC} = \sum_{j=1}^{p-s} h_j^{-1}\boldsymbol{t}_j^{\intercal}\boldsymbol{X}^{\intercal}\boldsymbol{y}\boldsymbol{t}_j \tag{2.30}$$

where $\boldsymbol{T}$ is the $p \times p$ orthogonal matrix whose columns are the eigenvectors $\boldsymbol{t}_j$ corresponding to the eigenvalues $h_1, h_2, \ldots, h_p$ from $\mathbf{X}^\intercal\mathbf{X}$. We also have that $\boldsymbol{Z} = \boldsymbol{XT}$ and that $\boldsymbol{\Lambda} = \text{diag}(h_1, h_2, ..., h_p)$. We now define $\hat{\boldsymbol{\alpha}} = (\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{y}$ and from that it follows that $\boldsymbol{T}\hat{\boldsymbol{\alpha}}_{PC}$ is defined as:

$$\boldsymbol{T}\hat{\boldsymbol{\alpha}}_{PC} = \begin{bmatrix} \hat{\boldsymbol{\alpha}}_1 \\ \hat{\boldsymbol{\alpha}}_2 \\ \vdots \\ \hat{\boldsymbol{\alpha}}_{p-s} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ Including the } p-s \text{ first components from } \hat{\boldsymbol{\alpha}}$$

So equation (2.30) can be seen as a regression with the first $p - s$ principal components as the independent variables/regressors. The principal components are orthogonal so one can just add the univariate regression results with one principal component as the regressor for the others [17].

2.5 Robust regression methods

In this section we will present two robust regression methods; Least absolute deviation regression and M-estimation regression. These robust regression methods handle outlier values in the data better than the OLS regression method.

2.5.1 Least absolute deviation (LAD) regression

The LAD (Least Absolute Deviation) regression method minimizes the residual sum of absolute errors [7].

$$\hat{\boldsymbol{\beta}}_{LAD} = \underset{\boldsymbol{\beta}}{\text{argmin}}\{\sum_{i=1}^{N}|y_i - \beta_0 - \sum_{j}^{p}\beta_j x_{ij}|\} \tag{2.31}$$

LAD regression is computationally expensive with data sets containing many data points n as it needs to be solved using an iterative process, on the other hand it is more robust than the OLS regression as outlier points do not affect the regression model to the same extent. LAD regression will not be performed in this thesis however it is a good introduction to the Huber form of the M-estimation which is described next.

2.5.2 M-estimation (Huber regression)

The M-estimation regression method minimizes the residual sum of a specific error function $\rho(e)$. One can therefore observe that OLS regression and LAD regression are special cases of M-estimators where the error function are $\rho(e) = e^2$ and $\rho(e) = |e|$ respectively [7][14].

$$\hat{\boldsymbol{\beta}}_M = \underset{\boldsymbol{\beta}}{\text{argmin}}\{\sum_{i=1}^{N}\rho(y_i - \beta_0 - \sum_{j}^{p}\beta_j x_{ij})\} \tag{2.32}$$

A good choice of an error function $\rho(e)$ is one that meets the following properties [21]:

- The error function is allays non-negative, $\rho(e) \geq 0$
- The error function should be equal to zero when the error is zero, $\rho(0) = 0$
- The error function should be symmetric, $\rho(e) = \rho(-e)$

- The error function should be monotone for the absolute value of the errors, $\rho(|e_1|) \geq \rho(|e_2|)$ when $|e_1| \geq |e_2|$

We now look closer at a specific choice of $\rho(e)$ which is the Huber M-estimate which is given by the following formula:

$$\rho(e) = \begin{cases} e^2 & \text{if } -k \leq e \leq k \\ 2k|e| - k^2 & \text{if } e < -k \text{ or } k < e \end{cases} \tag{2.33}$$

So the Huber M-estimator combines the best properties of OLS and LAD estimation and we also see that the error function in equation (2.33) meets the four properties from above. The Huber regression method is more robust than the OLS method as the residual has the same behaviour as the LAD method for large errors which is favourable as the LAD method is less sensitive to outliers than the OLS method. The parameters in the Huber regression are chosen in a way so the error function $\rho(e)$ is continuous. Huber recommends a k-value of $1.5\hat{\sigma}$ where $\hat{\sigma}$ is an estimate of the standard deviation σ of the population of random errors and this recommendation will be used in this report [7].

The solution to the Huber regression method from equation (2.32) with the error function from equation (2.33) can not be written in a closed form as for the OLS regression method in equation (2.5). One therefore needs to use an algorithm to find the solution to equation (2.32), a commonly used algorithm is the *iteratively reweighted least-squares* (IRLS) (see [21] for a description of this algorithm) [14][21].

3 Current state of knowledge / Literature review

Here we will present the current state of research when it comes to hedonic index modeling. The section will mainly be structured around the Eurostat Handbook on Residential Property Indexes [11] but will also cover some research papers and a description of the HOX-index (a residential property index that already exists for the Swedish market).

3.1 Handbook on Residential Property Prices Indices (RPPIs)

The statistical office of the European Union (Eurostat) has published an extensive guide on how to model Residential Property Price indexes [11]. In this guide the authors describe four different methods that are commonly used in modeling of RPPIs. The main methods are:

"Stratification" of the transactions according to some characteristic of the property that was sold. One then creates different cells and takes the average price in each cell and then uses these average prices to model the residential property index. A stratification method with only one cell becomes a pure average price index.

In the *"repeat sales index method"* the quality mix problem is handled by calculating the index from objects that have been sold in both the base period and the period for which one is interested in modeling the index. One then assumes that the quality of the object is the same in both periods and the repeat sales is later used to model an index.

"The hedonic regression model approach" is data intensive however it takes the changes in the quality of the objects into account. In this method a linear pricing model is built for the objects then this model is used to model the index either using a time-dummy approach or a imputation approach. This method is described in more detail in the methodology section and this thesis will primary focus on this type of index models.

The fourth method described is *"the assessment based pricing method"* which takes the tax valuation of the property into account when valuing the property for modeling of the index.

The handbook describes the four methods mentioned above in depth in one chapter each. The handbook also discusses uses for RPPIs and how one could collect data to model RPPIs.

3.2 Research papers and related Books

The modeling of indices for heterogeneous goods is an important field of study and there exist many theoretical and empirical articles about this subject. We will describe the core results from some of them that are relevant to this thesis.

In the paper "Price and quality of desktop and mobile personal computers: A quarter- century historical overview" (2001) the authors examine the price development of personal computers in the period 1976-1999. The paper compares different ways to model the price increase and finds that the model results are sensitive to the underlying change of the characteristics [1] . The characteristics of a personal computer changes much faster than those for a Swedish apartment but one should still have this in consideration when modeling RPPIs.

In the paper " Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and Other Approaches" (2010) the author discusses the main methods currently used for modeling price indexes [8]. The author states that the time dummy method is more restrictive than the double imputation index method but the time dummy method can be useful if the data is sparse as it preserves the degrees of freedom in the regression model. The author also states that the double imputation method is

preferable over the single imputation method if the index is modeled for unmatched items.

The book "Price Index concepts and Measurement" (2009) by W. Erwin Diewert, John S. Greenlees and Charles R. Hulten discusses index modeling methodologies and recent research papers relating to index modeling. Chapter 4 in this book discusses the differences between an hedonic imputation approach and a time dummy approach in modeling indexes. We will now look closer at this chapter and summarize the key finding which is interesting for this report [20]. This chapter has the same authors and much similarities with the article "Hedonic Imputation Versus Time Dummy Hedonic Indexes" [9] and therefore only the book is described here. The authors show that the hedonic imputation method and the time dummy method produce identical indexes if the average characteristic are constant in all periods. This means that the average values of the independent variables should be constant in all time periods [20].

3.3 The HOX-index

There exists a residential property index in Sweden, the "Nasdaq OMX Valueguard-KTH Housing Index" that has the ticker HOX. The HOX-index is a hedonic time dummy index and a constant quality index. The index is modeled using a weighed least squares method. The purpose of the index is to measure the price development of a typical one family house, apartment or a combination of the two. The HOX-index is based on sales transaction data from Swedish real estate brokers and excludes the sales of newly constructed property [18].

The HOX-index has the index base month of January 2005 and has monthly index values from that base date. As mentioned earlier the HOX-index is modeled by using a time dummy hedonic pricing method, this method has the drawback that earlier index values could change when more and newer data is added to the model. This property is not favourable for an index (especially not for one that is used on an exchange, which is the case for the HOX-index). This problem is solved by only adding the newest index point to the index when the index is updated with new monthly data [18].

The HOX-index is modeled using the following regression model:

$$\log(y_i^t) = \beta_0 + \sum_{\tau=1}^{T} \delta^\tau D_n^\tau + \sum_{k=1}^{K} \beta_k x_n^t + \epsilon_n^t \tag{3.1}$$

where y_i^t is the price of the property, δ^τ the time dummy coefficient that creates the index, D_n^τ a dummy variable that indicates if the property was sold in the specific month, x_n^t the descriptive variables used in the index like size of the property and β_k is the parameter vector for the different properties. The index contains T periods, and therefore T index points, and is constructed using K characteristics [19].

The HOX-index has some parameters in the model that handles the geographical position of the property, the distance to the centre of the city is included and represents a price gradient. The city for where the index is calculated is also split into 4 quadrants (northwest, northeast, southwest and southeast)[19].

The HOX-index has handled the problem with outliers and measurement errors in modeling of the index by using the following three step procedure [19]:

1. Remove outliers using a Cook's distance test.
2. Do a robust regression of the data using Huber regression and biweighting.
3. Use an iterative cross validation approach to test the model.

One can se the HOX-index for apartments in Stockholm and Sweden in general during the period from 2005 to 2016 in figure (3.1) [18].

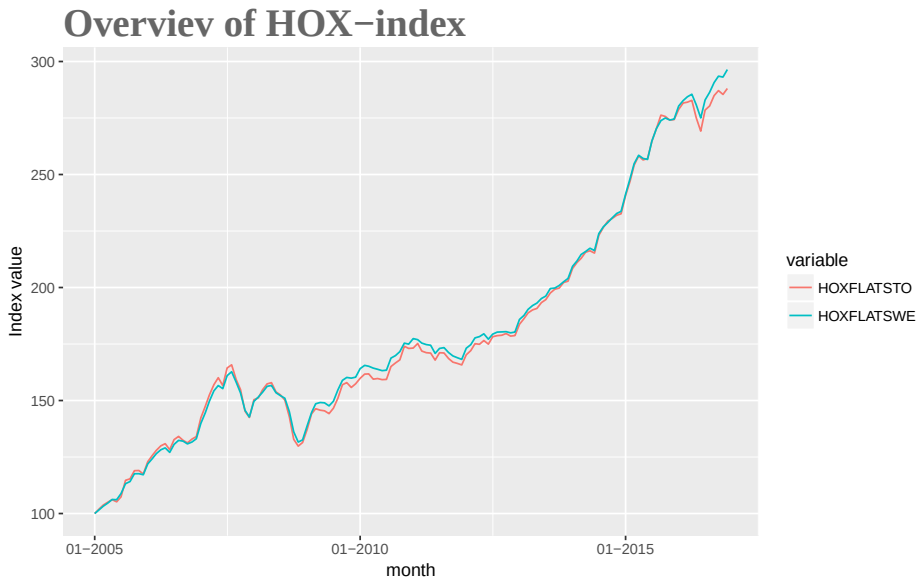


Figure 3.1: An overview of the HOX-index for apartments in Stockholm (the line HOXFLATSTO) and in Sweden (the line HOXFLATSWE).

4 Data overview and pre-processing

The data used in the report will be described in this section. We will first describe how the data were collected, which parameters are included in the dataset, how the geographical data is handled/transformed and an overview of the missing data in the dataset.

4.1 Data collection

The data used in this thesis is provided by Booli Search Technologies AB (called Booli in the rest of the thesis) and is downloaded by Booli's web API. Booli has created their database of the data by collecting the data from the different real estate agencies web pages with help of a web crawler (a program that search the web for information). The data that is collected is primarily from objects that were not sold before the "screening" of the objects which means that not all sold objects in Sweden are included, however the data set contains a majority of the sold objects. The fact that Booli collects the data with a web crawler makes the data second source data and all real estate agencies do not provide all the data points that we are interested in so this creates a problem with missing data.

4.2 Variable overview

In this section we will list the data from the data set that will be used to model the index and show how many data points there exists for the different geographies and for the different time periods considered. There exists different data points for the different types of listings but we will focus on the apartment category.

4.2.1 Apartment data variables

Here we present at the data variables that can be useful for a valuation model for the apartment category in Stockholm.

soldPrice	The price the apartment was sold for in SEK
rent	The rent of the apartment in SEK/month
floor	The floor of the apartment
livingArea	The living area of the apartment i m^2
rooms	The number of rooms in the apartment
constructionYear	The year the apartment building was constructed
objectType	The type of object, in this case all are "Lägenhet", i.e apartment
operatingCost	The cost of operating the apartment SEK/year
soldDate	The date tha apartment was sold
isNewConstruction	A dummy variable to see if the apartment is newly built (sold for the first time)
location.position.latitude	The latitude position of the apartment
location.position.longitude	The longitude position of the apartment
location.region.municipalityName	The municipality the apartment is located in
location.region.countyName	The county that the apartment is located in
location.distance.ocean	The distance to the ocean in meters

location.distance.water The distance to the nearest water body in meters

Other variables are also available including some internal ID-variables that are not relevant for a valuation model and these are therefore excluded. We create dummy variables for the constructionYear variable due to previous analysis done by Booli (see table (4.1)). We have the following creation of dummy variables which also provides a solution for the missing data problem for the constructionYear variable. The dummy variable gammal.CT.dummy will be left out of the regression model as it is set to be the base case.

Dummy name	lower boundary $>$	upper boundary $\leq$
gammal.CT.dummy	0	1934
funkis.CT.dummy	1934	1958
folkhem.CT.dummy	1958	1965
miljonprogram.CT.dummy	1965	1975
osubventionerat.CT.dummy	1975	1994
modern.CT.dummy	1994	2010
nyproduktion.CT.dummy	2010	9999
missing.CT.dummy	???	???

Table 4.1: The construction of a dummy family for the constructionYear variable. Every data point that had a missing value for constructionYear was assigned to the missing.CT.dummy category. The dummy variable takes on the value 1 if it is in the time interval and 0 otherwise.

4.3 Monthly data generating process

We are also interested in the number of data points for the different time periods (monthly time periods), figure (4.1) shows that the number of sold apartments differ substantially between the cities Stockholm, Uppsala and Eskilstuna. We also see that the number of sold apartments show a very cyclical pattern with many apartments sold during the spring and autumn.

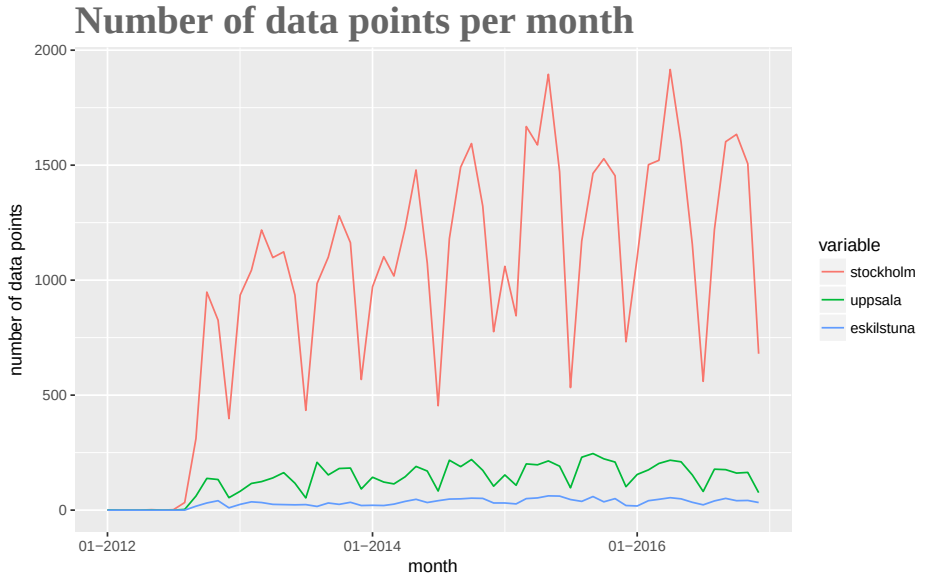


Figure 4.1: The number of available data points for apartments per month for the different cities during the period of January 2013 to December 2016. One can clearly see that the number of sold apartments exhibits a cyclical behaviour in all the cities.

4.4 Geographical data

Geographical coordinates should not be put directly into a linear regression model as they probably do not have a linear impact on the dependent variable. This problem is solved by assigning every data point to a geographical area that was constructed by combining adjacent postal codes by the clustering algorithm called Skater. In this algorithm similar postal codes are grouped together to form bigger geographical areas which are given an area code (the area code is a number but this number is irrelevant for the analysis and is only used to separate the different areas). The clusters and geographical split were provided by Booli. The packages sp and rgdal in R were then used to assign a number to every geographical area by using location.position.longitude and location.position.latitude. One dummy variable was then created for each geographical area, the dummy taking the value 1 if the object was in the specific area and the value 0 otherwise. There exists 50 geographical areas which contain at least 5 sold objects during the time period from January 2013 to December 2016.

The construction of dummy variables from geographical coordinates solves the problem of including coordinates in a regression model. We will also use the distance to the nearest ocean (location.distance.ocean) and the distance to the nearest water body (location.distance.water) as representations of the geographical position in our valuation models.

4.5 Missing data

We would like to know how severe the problem with missing data is for the Stockholm data set. The variables that could be useful for the valuation model are listed in table (4.2) with the total number of missing data points for the different categories. We see that the number of missing variables differs quite substantially over the different variables. We now plot the number of missing variables per month to see how the missing data is distributed and the result is presented in Figure (4.2).

	nr of missing values	% of total
rent	438.00	0.01
livingArea	152.00	0.00
rooms	116.00	0.00
floor	11295.00	0.20
location.distance.ocean	7043.00	0.12
location.distance.water	35.00	0.00
constructionYear	7084.00	0.12

Table 4.2: The number of missing data points for the Stockholm data during the period of January 2013 to December 2016

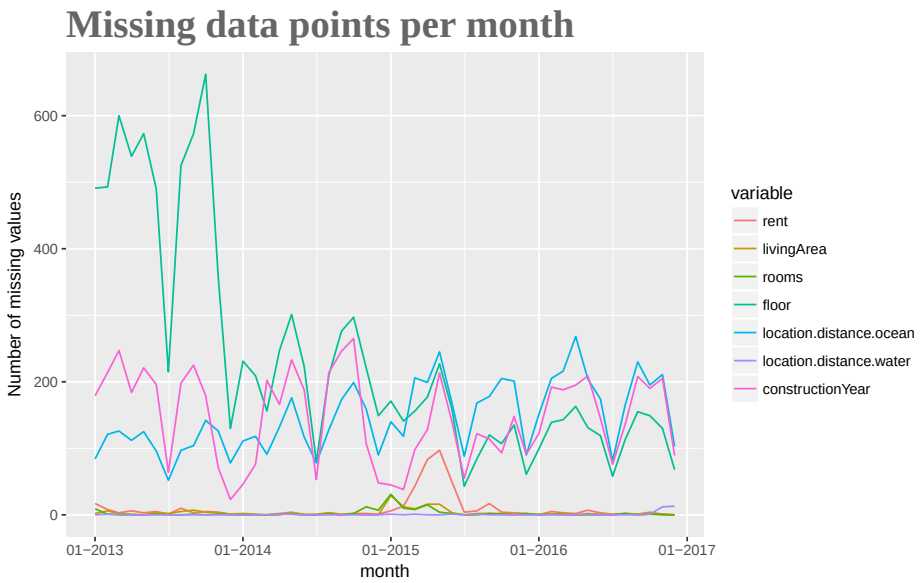


Figure 4.2: The number of data points missing per month in the Stockholm data for the time period January 2013 to December 2016. We can clearly se that the variables floor, location.distance.ocean and constructionYear is the variables that have several data points missing.

From figure (4.2) we see that the number of missing data points has a similar distribution as the number of sold objects from figure (4.1) with some exceptions like floor data missing in the start of the examined period and rent data missing in the beginning of 2015. Missing data can lead to a biased valuation model and therefore biased RPPIs so we will later address how we can avoid this problem with missing data.

5 Methods

In this section we will present the methods and methodology that are going to be used for the modeling and analysis of the RPPIs. We will also introduce some properties of a linear model that is useful for us in the valuation model before we describe ways to validate the RPPIs.

5.1 Hedonic time dummy variable model

The first of the two main hedonic index estimation models that will be examined in this report is the hedonic time dummy variable model. The main idea behind this model is to create a linear regression model that includes a dummy variable for each of the time periods that should be included in the index (except the period which becomes the base period). One can then extract the RPPI from the coefficients of the dummy variables in the model. One most often use a log-linear model when modeling a time dummy variable hedonic index model:

$$\log(p_n^t) = \beta_0 + \sum_{\tau=1}^{T} \delta^\tau D_n^\tau + \sum_{k=1}^{K} \beta_k z_{nk}^t + \epsilon_n^t \tag{5.1}$$

where the time dummy variable D_n^τ has the value 1 if the observation comes from period τ and 0 otherwise. The variable δ^τ is the parameter term that describes the time period dependence on the dependent variable. In the time dummy variable model z_{nk}^t has the same meaning as in the linear model, which means that it is the value of "quality" k for property n. The index contains T periods, and therefore T index points, and is modeled using K characteristics.

From the model in equation (5.1) the time dummy index going from period 0 to period t is given by [11]:

$$P_{TD}^{0t} = exp(\hat{\delta}^t) \tag{5.2}$$

where $\hat{\delta}^t$ is taken from equation (5.1).

One could rewrite equation (5.2) if OLS regression is used for fitting of the linear model. [11]

$$P_{TD}^{0t} = \frac{\prod_{n \in S(t)} (p_n^t)^{1/N(t)}}{\prod_{n \in S(0)} (p_n^0)^{1/N(0)}} exp[\sum_{k=1}^{K} \hat{\beta}_k (\bar{z}_k^0 - \bar{z}_k^t)] \tag{5.3}$$

This alternative form of the formula shows why one does not need to account for the bias correction term of $s^2/2$ for transformation from a log-linear, as the correction term is the same for both the nominator and the denominator as the error term is constant for the whole regression model.

A benefit of the time dummy model is that pooling the data from different time periods together preserves the degrees of freedom in the regression model which leads to smaller standard deviations. The cost of this is that one makes the assumption that the coefficients would be constant over time (which probably is not the case) [11].

5.2 Characteristics Prices Approach

The other main index modeling approach that is going to be examined in this report is the hedonic characteristic/imputation approach. The main idea behind this approach is to create a valuation model per period of the index by using the data from this time period. The index is then modeled by using the same input in the both time periods and divide the price obtained in the end period model by the price obtained by the start period model. The index from the base period 0 to the end period t is modeled using the Characteristics Prices Approach using the following formula:

$$P_{CP}^{0t} = \frac{\hat{p}^t}{\hat{p}^0} = \frac{\hat{\beta}_0^t + \sum_{k=1}^{K} \hat{\beta}_k^t z_k^*}{\hat{\beta}_0^0 + \sum_{k=1}^{K} \hat{\beta}_k^0 z_k^*} \tag{5.4}$$

The main idea is to compare the modeled price using a linear model fitted using data from different time periods. Different choices of the "quality" parameter z_k^* will give rise to different indexes, and some standard choices will be described next [11].

5.2.1 Laspeyres-type

The two main approaches for the quality parameter in equation (5.4) is to use the mean values of the characteristics from the base period or from the end period. By setting the quality parameters to the mean values of the base period $z_k^* = \bar{z}_k^0$ in equation (5.4) we get the Laspeyres-type characteristics price index [11]. This give us the following formula for the index:

$$P_{CPL}^{0t} = \frac{\hat{\beta}_0^t + \sum_{k=1}^K \hat{\beta}_k^t \bar{z}_k^0}{\hat{\beta}_0^0 + \sum_{k=1}^K \hat{\beta}_k^0 \bar{z}_k^0} \tag{5.5}$$

5.2.2 Paasche-type

By using the mean values from the end period instead $z_k^* = \bar{z}_k^t$ in equation (5.4) we get the Paasche-type characteristics price index [11]:

$$P_{CPP}^{0t} = \frac{\hat{\beta}_0^t + \sum_{k=1}^K \hat{\beta}_k^t \bar{z}_k^t}{\hat{\beta}_0^0 + \sum_{k=1}^K \hat{\beta}_k^0 \bar{z}_k^t} \tag{5.6}$$

5.2.3 Fisher-type

We get the Fisher type characteristic price index by taking the geometric mean of the Laspeyres-type index and the Paasche-type index (equation (5.5) and equation (5.6)) [11]:

$$P_{CPF}^{0t} = [P_{CPL}^{0t} P_{CPP}^{0t}]^{1/2} \tag{5.7}$$

The reasoning by using the Fisher type index is to cancel the effects of under/over pricing that a large change in the characteristics between the start and end period can lead to. The Fisher index will therefore be used to model the indexes in this report as it is the most robust alternative of the three methods mentioned above.

5.3 Hedonic Imputation Approach

The hedonic imputation approach is a proposed solution of the problem that one can not observe the period t prices of the properties sold at period 0 because most of the properties will not be resold during period t. The same problem is that one can not observe the period 0 price for properties sold at period t if the property was not sold in both periods (which is a rare case). To circumvent this problem one can impute (use a valuation model for approximation) the prices that are missing.

5.3.1 Single Imputation approach

In the single imputation approach the observed prices are left unchanged so in the Laspeyres case the prices in period 0 are left unchanged and the prices for period t are modeled. The index for the single Lespeyres imputation approach is the following:

$$P_{HI}^{0t} = \frac{\sum_{n \in S(0)} 1 * \hat{p}_n^t(0)}{\sum_{n \in S(0)} 1 * p_n^0} \tag{5.8}$$

5.3.2 Double Imputation approach

In the double imputation approach both the prices for period 0 and period p are modeled. The index for the double Laspeyres imputation approach is the following:

$$P_{HI}^{0t} = \frac{\sum_{n \in S(0)} 1 * \hat{p}_n^t(0)}{\sum_{n \in S(0)} 1 * \hat{p}_n^0} \tag{5.9}$$

5.3.3 Results for OLS

For the double imputation approach the arithmetic imputation indexes become the same as for the characteristics prices approach in equations (5.5), (5.6) and (5.7), independent of the use of model (as long as the model is linear) [11]. If one uses an OLS model the single imputation approach will be the same as (5.5), (5.6) and (5.7) [11]. We will therefore refer to the name double imputation approach when we talk about the models in equation (5.5), (5.6) and (5.7) in the remaining of the paper. Because double imputation method is the more general name of the methodology and the characteristics approach is a special case of the imputation approach.

5.4 Overview of index modeling

The valuation method in the hedonic double imputation index approach above is that one creates two valuation models, one for period 0 (the base period) and one for the period t (the end period). One then uses the same input values in both models and compute the quota of the two predicted values. Therefore to model a good hedonic double imputation index one should create accurate and robust valuation models for the residential property. The model creation will be done by analysing the December 2016 sales of the apartments in Stockholm to create a valuation model that can be used in the imputation/characteristic approach.

To model a good time dummy index one would like to create a good valuation model for the whole time period where the focus is extra high on the time dummy parameters. We would like this model to be as robust as possible with small multicollinearity for the time dummy parameters.

5.5 Creating a valuation model

A linear valuation model is first created for the Stockholm region to see which variables lead to the best prediction model. We will use an OLS approach when the model is first constructed and we will then see if the model needs improving, possible improvements are:

- Transformation of the variables in the model.
- More advanced versions of linear regression (Huber regression, ridge regression, lasso regression and PCR).

We will create a linear model for December 2016 and use this model as a benchmark to decide the format of the linear model used in the hedonic imputation approach. We are going to take a similar approach as the HOX-index, when we create our linear valuation model for December 2016.

1. Create and investigate a base model with all the data and an OLS approach.
2. Perform an all possible regression analysis to select the best linear model.
3. Use the cross validation approach to select the best parameter set for the model (including different combinations of cross terms).
4. Investigate different methods to detect high leverage and influence points.
5. Investigate different transformations of the dependent and independent variables.

6. Investigate different methods to deal with missing data.

7. Investigate the residuals and see if any robust regression method is needed for the modeling of the index and compare different robust regression methods.

8. Implement the different models from above for the different index approaches and use cross validation to see which index that performs the best.

5.6 Modelling of the RPPIs

We will use the best linear valuation model from the analysis of the 2016 December data to model an arithmetic hedonic imputation RPPI. We will use the Fisher index method from equation (5.7) for the modeling of the arithmetic hedonic imputation RPPIs. We will model indexes by this approach by using 3 different models, one with the same valuation equation in all periods, one with a dynamic choice of the location dummy variables and one with dynamic choice of the location dummy variables and a sliding modeling of the index. These models are described in more detail below.

5.6.1 Arithmetic hedonic imputation method (model 1)

We will model a index where the same model is used in every period where, the base period January 2013 is used as the denominator in the index equation (5.5) - (5.7). The linear model is chosen by the analysis of the December 2016 data by the method described above.

5.6.2 Arithmetic hedonic imputation method(model 2)

The same as model 1 however here the cluster dummy only needs to have 5 data points in the cluster in the specified time period for the index point and base point. This leads to an inclusion of a greater number of clusters.

5.6.3 Arithmetic hedonic imputation method(model 3)

This model uses the same choice of location dummy variables as (model 2) but in this model the index is modeled by modeling an index for every period i in the denominator and $i+1$ in the nominator and then multiplying the one period indexes to get an index for the base period to the end period. We describe this by the following formula:

$$P^{0t}_{HI(model3)} = P^{01}_{HI} P^{12}_{HI} \dots P^{t-1t}_{HI} \tag{5.10}$$

where $P^{01}_{HI}, P^{12}_{HI}, \dots$ is modeled by the same method as (model 2). A potential problem with model 3 is the problem of error propagation which means that a error in one of the earlier periods will be propagated and increase the error in the later periods.

5.6.4 Hedonic Time dummy method

The hedonic time dummy index is created by creating the best version of equation (5.1) by performing an all subset analysis and a cross validation analysis of the model. The index values are then modeled using equation (5.2).

5.6.5 Average method

An average method index is modeled by taking the average value of the sold price in the base and end periods and divide the average price in the end period with the average price in the base period:

$$P^{0t}_{Average} = \frac{\bar{p}^t}{\bar{p}^0} \tag{5.11}$$

The average approach index is modeled for the whole non-cleaned data set and for the cleaned data set with missing values removed which is used to model the indexes with the other methods.

5.7 Validation of the model

The index modeling methods described above creates a point estimate for each index value, this point estimate does not say anything about how good the index modeling method is so we shall here describe the different methods to validate the model that are used in the report. Methods to validate both the individual valuation model for one time period and validation methods for the RPPIs will be described.

5.7.1 Cross validation

In the first step of the model creation the model is created using the whole data set. The problem with this approach is that it can create a problem of over fitting and this problem can be handled using cross validation. We will use an n-fold cross validation approach where we will split the data set into n randomly assigned different parts and then fit the model to the union of $n-1$ of the data sets and then calculate the error with the subset that was not used in the model fitting. We will then use the same data sets and do this procedure so all of the n data sets have been the testing set and then add all the errors. We can then compare different models with a lower risk of over fitting the model to the current data. So the algorithm becomes [12]:

Data: the whole data set N and the model to test
Result: A model error (of the sort one is interested in)
Randomly split the data set in n equally sized parts;
for *i in 1 to n* **do**
    Fitt/train the model with the data set N - the i:th data set;
    Calculate the error by using the test set i;
    Add the actual error to the total error;
end

Algorithm 1: Performs the cross-validation for validation of the valuation models. Different metrics will be calculated in the cross validation including SS.mean which is the mean squared error , AS.mean which is the mean absolute error, R^2 which is an R^2-value calculated for the test set and RA which is a version of R^2 where the squared error is change to absolute error.

The cross validation will only be performed to test the regression models as the data needed to do a cross validation of the modeled indexes does not exist, as the data does not contain price information for the same object in two different time periods.

5.7.2 Plotting the coefficients for the different time periods

One validation approach that can be used for hedonic imputation RPPIs is to plot the coefficients for the different time periods that the index is modeled for. This graph should then be as flat as possible as this is a sign of a robust valuation model, a model where the coefficients change much between the time periods would be a sign that the model suffers from high multicollinearity because the underlying price factors should not change much from month to month.

This is not a quantitative validation model but a qualitative one and it has its limitations. The coefficient graphs would be flatter if the model contained fewer variables even if the all subset regression analysis showed that many variables should be used. However this can be used to compare RPPIs with the same number of independent variables that were created with different regression methods, for example a comparison between OLS and Huber regression.

5.7.3 Plotting the characteristics for the different periods

Plotting the mean value of the characteristics for each period is helpful in deciding if one should use a time dummy or an hednonic imputation approach for modeling the RPPI. This notion comes from that the time dummy approach and hedonic imputation approach is equivalent if the characteristics are constant in all periods, the methods are similar if the characteristics do not deviate much between

periods. (See the theory section and [20] for a more thorough explanation). Cross validation can not be used to test the index performance so this analysis will therefore be an important component in deciding between the time dummy model and the double imputation model.

5.7.4 Bootstrap

Another validation approach that is used is the Bootstrap method which is based on a re-sampling of the available data. One version of the non-parametric bootstrap version for calculating the MSE for a parameter θ is:

$$d(\theta^*) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\theta^*_{i,obs} - \theta^*_{mean,obs})^2} \tag{5.12}$$

where $\theta^*_{i,obs}$ is a realisation of the parameter from sampling nr i, B is the number of re-samples that is made and $\theta^*_{mean,obs} = \sum_{i=1}^{B} \theta^*_{i,obs}/B$ is the arithmetic mean [2].

Data: the whole data set Ω and the index methodology to test
Result: n calculated versions of the index
for *i in 1 to n* **do**
    Draw N data points with replacement from the whole dataset Ω;
    Model the index by using the bootstrapped dataset;
end

Algorithm 2: Perform the index modeling using the bootstrap algorithm. The results will then be used to calculate both the MSE for each index and time period and for creating empirical confidence intervals for the indexes.

The bootstrap method from above will be used to calculate the MSE for the different index methods. The MSE will be smaller than the real value as it contains many doublets and triplets of the same data points, but this analysis will help to get a sense of the relative errors between the methods. We will also use the bootstrap sampling to model empirical confidence intervals for the indexes.

The bootstrap method will also be used to model an 95% empirical confidence interval for the modeled indexes. The confidence interval is modeled by taking the 2.5% and 97.5% empirical quantile for every index period.

6 Results and Analysis

In this section all the analysis described in the Methods section is implemented and analysed. It starts with an analysis of the December 2016 valuation model and uses the results from that analysis to model the hedonic double imputation indexes. The time dummy index is then modeled using similar analysis that was used for the December 2016 valuation model. The indexes are then evaluated using both qualitative and quantitative methods.

6.1 Creating the linear model

In this section we will describe the procedure when the linear valuation model for December 2016 is created. We will try to find the "best" model using the analysis described in the method section. We first consider the following model:

$$\begin{aligned}\mathsf{soldPrice}_t = \beta_0 &+ \beta_1 \mathsf{rent}_t \\ &+ \beta_2 \mathsf{livingArea}_t \\ &+ \beta_3 \mathsf{rooms}_t \\ &+ \beta_4 \mathsf{floor}_t \\ &+ \beta_5 \mathsf{location.distance.ocean}_t \\ &+ \beta_6 \mathsf{location.distance.water}_t \\ &+ \beta_7 \mathsf{funkis.CT.dummy}_t \\ &+ \beta_8 \mathsf{folkhem.CT.dummy}_t \\ &+ \beta_9 \mathsf{miljonprogram.CT.dummy}_t \\ &+ \beta_{10} \mathsf{osubventionerat.CT.dummy}_t \\ &+ \beta_{11} \mathsf{modern.CT.dummy}_t \\ &+ \beta_{12} \mathsf{nyproduktion.CT.dummy}_t \\ &+ \beta_{13} \mathsf{missing.CT.dummy}_t + \epsilon\end{aligned} \tag{6.1}$$

In this model we have set the dummy gammal.CT.dummy as the base value for the constructionYear dummy family of dummy variables to avoid perfect collinearity.

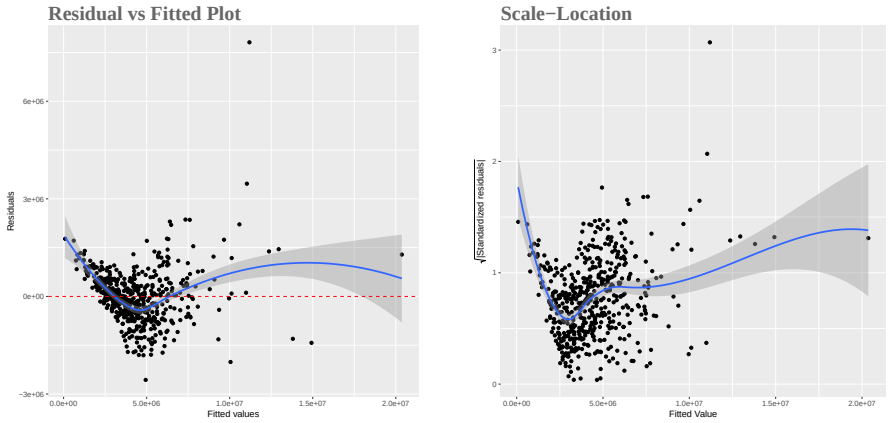
From this model we get a result with 161 of 677 values deleted due to any missing value being "NA" for the data point. We would therefore like to examine the number of missing values of the data and address the issue. The number of missing values for each independent variable is shown in table (6.1) and we can see that the problem with missing variables in the constructionYear variable is handled by creating a dummy missing.CT.dummy that indicates a missing value of the variable construciton Year.

We remove all the rows with one missing value in one of these parameters in the initial data analysis to get the same data set in all the regression models so it will be easier to compare the models. This is just a fix for the initial data-analysis, later we will investigate how the missing data points could be handled.

We now analyse the residuals for the first model (with the data with missing values removed) shown in equation (6.1), see figure (6.1). We can clearly see in the figure that the residuals for the model are not constant for the different values of the fitted value, this is a problem with the model and should be addressed by trying other models.

	nr of missing values
rent	0.00
livingArea	0.00
rooms	0.00
floor	68.00
location.distance.ocean	103.00
location.distance.water	13.00
funkis.CT.dummy	0.00
folkhem.CT.dummy	0.00
miljonprogram.CT.dummy	0.00
osubventionerat.CT.dummy	0.00
mordern.CT.dummy	0.00
nyproduktion.CT.dummy	0.00
missing.CT.dummy	0.00

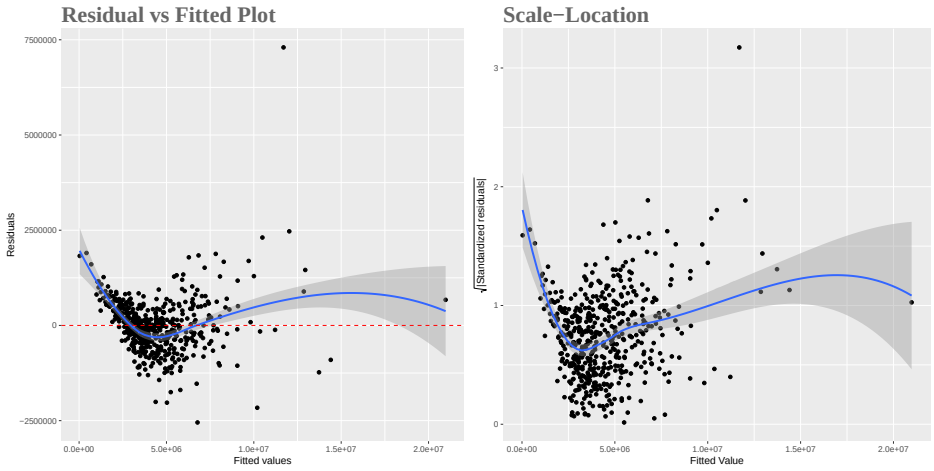
Table 6.1: The number of missing variables in the first model



(a) Residuals (b) Norm Residuals

Figure 6.1: The residuals equation (6.1) for the cleaned data set which does not contain any missing data. We can see that the residuals are not constant for the different fitted values.

We now analyse the first model and append the location cluster dummys to that data. We set the cut-off value for the cluster such that the cluster must contain at least 5 points to be used in the model. Figure (6.2) shows the residuals from equation (6.1) with all the clusters dummies added to the model (the base cluster becomes the clusters with too few data points to be included in the model). The model still exhibits residuals that are skewed which is a problem that could be handled by a different model formulation or by a transformation of the dependent variable (which will be done later in the report). However we shall try to find the best model with soldPrice as the dependent variable first.



(a) Residuals (b) Norm Residuals

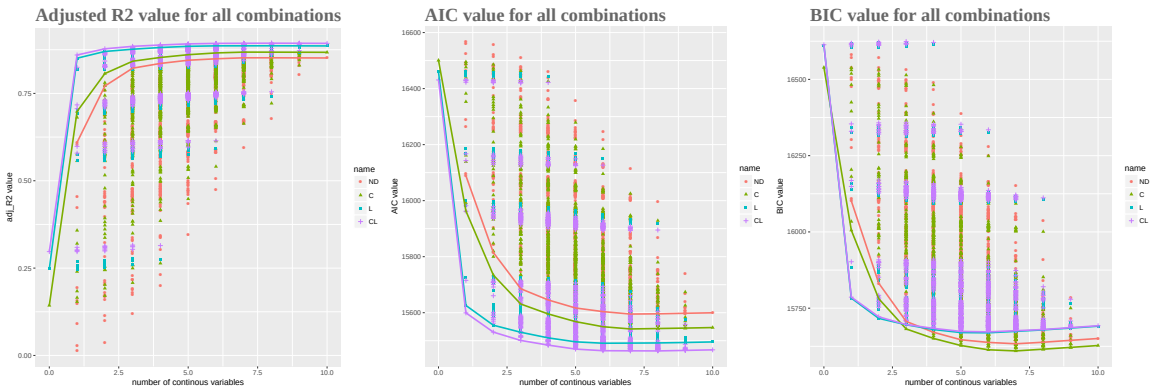
Figure 6.2: The residuals for equation (6.1) with the clusters dummies added to the model. The model is created with the cleaned data set which does not contain any missing data.

We now analyse the error terms plotted against the continuous variables in equation (6.1) with the cluster dummy variables added. This is done to illustrate the structure of the full model before an analysis of all possible models is performed. To do this we first plot the residual for the model vs these variables (for the non-dummy ones) which can be seen in figure (9.1) in the Appendix.

From figure (9.1) we note that the model contains one outlier point which has a residual of over $6 * 10^6$ We also see that we have some outlier points in the variables as well (for example an apartment with livingArea over 250 m^2). One can also see that error plots for Rent, Livingarea, Rooms and Floor have shapes which indicates that the model could be improved by including quadratic terms of these variables as well. We therefore add quadratic terms for Rent, Livingarea, Rooms and Floor in the all possible regression analysis.

6.1.1 All possible regressions

We have now performed some initial analysis on the model with the interesting parameters. In order to find a good regular linear model we perform a test to fit all the possible combinations of variables for the model, with square terms for the variables included as well. We do this with the cleaned data set and perform analysis on four different models which includes the following variables; the construction-year family, the location-dummy family, both the location-dummy and construction-dummy family and finally none of the dummy variables. The best results for the 4 different models will be marked with a line.



(a) Adj R2 (b) AIC (c) BIC

Figure 6.3: An all subset analysis for the linear model with the cleaned data. All points in the diagram represent a different combination of the subsets and the different shapes of points represent different dummy-families. ND stands for no dummies and is the regression model without any of the dummy families, C indicates that the construction time dummy family is included, L indicates that that the location dummy family is included and CL for both the location dummy and construction dummy families included. The optimal solutions for the different combinations of dummy-families are marked with a solid line.

We analyse all three graphs in figure (6.3) and the tables for the models with best values for adjusted R^2, sum of squared residuals, AIC and BIC. In the graphs we see that models with more regressors included in the model are favoured by adjusted R^2, sum of squared residuals and AIC while BIC favoureds the models with less regressors included. We now test the best models from the different methods using a cross validation approach (The 10 best models according to the different metrics can be seen in figure (9.3) and (9.4) in the Appendix).

6.1.2 Cross validation

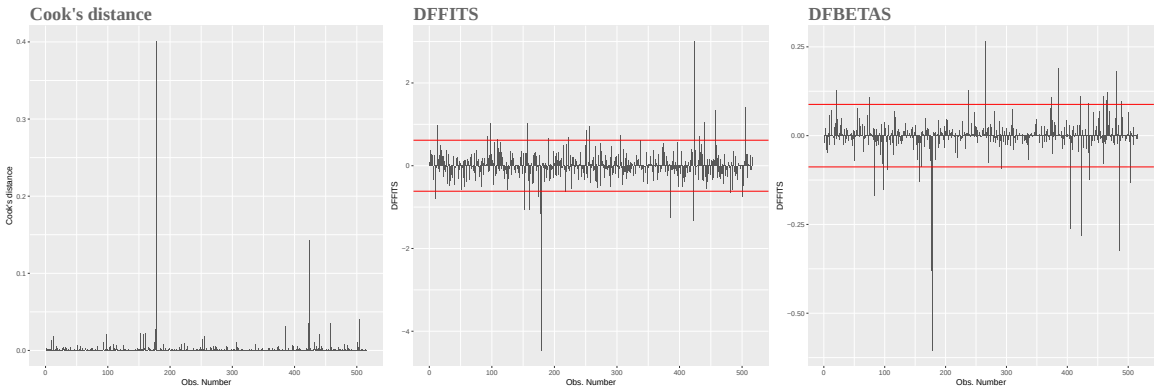
We now perform a cross validation on the three best models from the all subset analysis according to the adjusted R^2 condition, the AIC condition and the BIC condition (the best model according to the adjusted adjusted R^2 condition and the AIC condition are the same). We then analyse the R^2 value of the cross validated models and the result can be seen in table (6.2). The performance of the five first models are very similar but we choose model 4 for further analysis because this model has the best R^2 value.

model	SS.Mean	AS.Mean	R2	RA
1	$6.5745 * 10^{11}$	$5.4919 * 10^5$	0.8710	0.6518
2	$6.7748 * 10^{11}$	$5.5252 * 10^5$	0.8670	0.6497
3	$6.6558 * 10^{11}$	$5.5129 * 10^5$	0.8694	0.6505
4	$6.5109 * 10^{11}$	$5.5018 * 10^5$	0.8722	0.6512
5	$6.5734 * 10^{11}$	$5.4769 * 10^5$	0.8710	0.6528
6	$7.5595 * 10^{11}$	$6.0650 * 10^5$	0.8516	0.6155
7	$7.6815 * 10^{11}$	$6.0898 * 10^5$	0.8492	0.6139
8	$7.6321 * 10^{11}$	$6.0620 * 10^5$	0.8502	0.6157

Table 6.2: Cross validation performed on 8 different models where the first 3 models are the best models according to the adjusted R^2 measure from the all subset regression, model 4 and 5 are the second and third best according to the AIC criterion (The first model is the best according to the AIC criterion as well as to the adjusted R^2 criterion) and 6 to 8 are the three best according to the BIC criterion. The cross validation is performed with set.seed() in R which creates the same 10-fold cross validation split for all the models.

6.1.3 Influential point measure

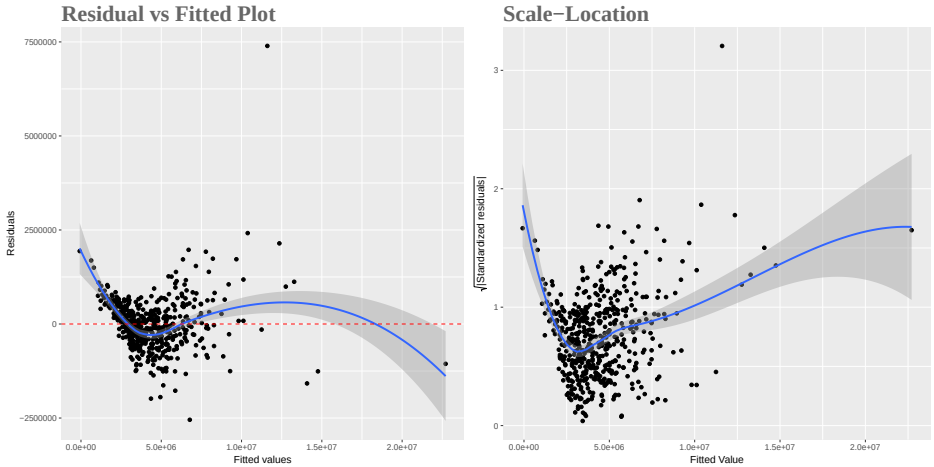
We analyse the different influence measures for the model described above and plot them with the cut-off value represented by the red horizontal line. The analysed influence measures are Cook's D with cut-off value cutoff $= 4/n$, DFBETAS with cut-off value cutoff $= 2/\sqrt{n}$ and DFFITS with cut-off value cutoff $= 2 * \sqrt{p/n}$. In figure (6.4) one can see that there exists many influential points and that a robust regression method might perform better than OLS regression.



(a) Cook's D (b) DFFITS (c) DFBETAS

Figure 6.4: The influencial measures Cook's D, DFFITS and DFBETAS plotted for model 4 from table (6.2) with the suggested cut-off values marked with a horizontal line. We can see that there are several values in every model that are larger than the suggested cut-off value indicating that the model contains high influence points.

We now investigate the residuals on this "best model". We see in equation (6.5) that the form of the residuals is not optimal and we shall therefore analyse a transformation of the model in the next section.



(a) Residuals (b) Norm Residuals

Figure 6.5: The residuals equation (6.2). The model is created with the cleaned data set which does not contain any missing data.

6.2 Transformation of the dependent variable

From the residual plots in the previous section we observe that the model exhibits a skewness so we perform a Box-Cox transformation of the dependent variable to see if that give us a better model. The Box-Cox Transformation is performed on equation (6.1) and the result is shown in figure (6.6).

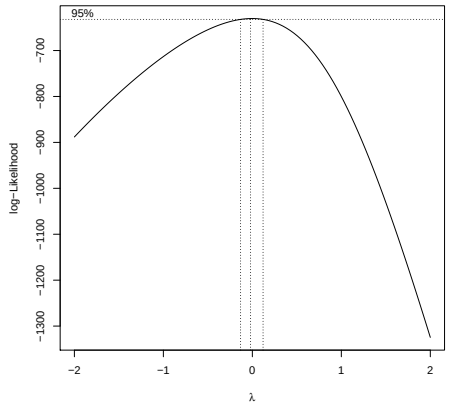
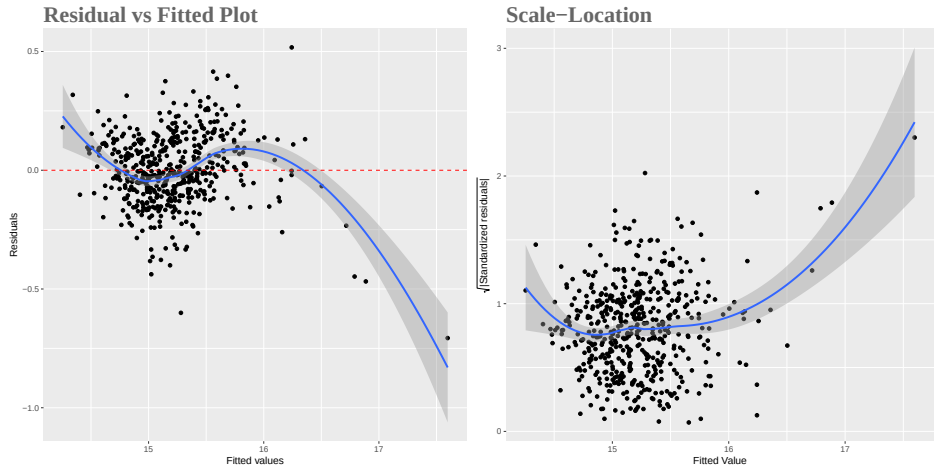


Figure 6.6: A Box-Cox transformation for the dependent variable for equation (6.1). The plot contains a 95% confidence interval for the optimal value of λ.

In figure (6.6) we see that a transformation value of $\lambda = 0$ which corresponds to taking the natural logarithm of the dependent variable soldPrice lies in the 95% confidence interval so we therefore perform a transformation of the dependent variable by taking log(soldPrice) and perform residual analysis on this model. We therefore begin with the initial model from equation (6.1) for the analysis of the transformed model. In figure (6.7) we see that the transformed model has less skewed residuals than the untransformed model but this model has some negative outliers for large values of the dependent variable. We see that the residuals are more evenly spaced with the exception of some outliers for the largest fitted values.



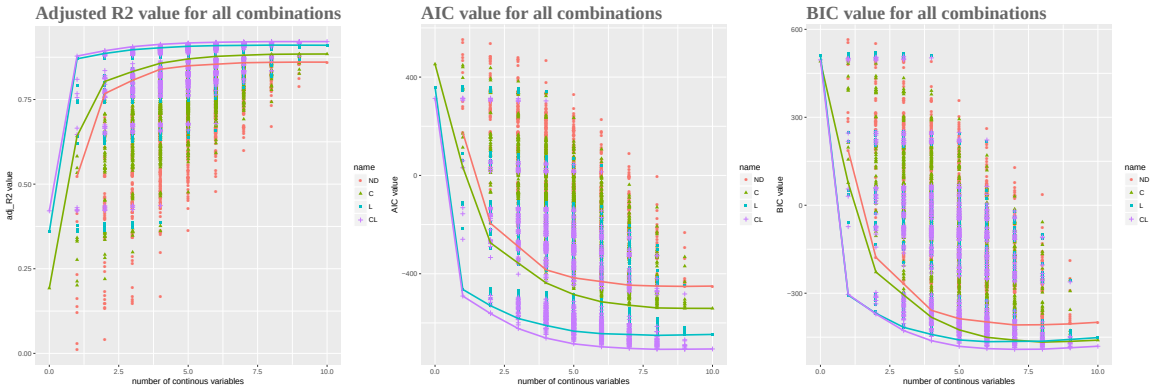
(a) Residuals (b) Norm Residuals

Figure 6.7: The residuals for equation (6.1) with the dependent variable soldPrice transformed with log(). The model is created with the cleaned data set which does not contain any missing data.

We now also analyse the residuals plotted against the continuous dependent variables which can be seen in figure (9.2) in the Appendix. We then perform a similar all subset analysis like we performed for the regular linear model to find the best log-linear model.

6.2.1 All subset analysis for loglinear model

In figure (9.2) in the Appendix we see that some quadratic terms could be of use so we perform an all subset analysis on the log-linear in the same way as for the linear model.



(a) Adj R2 (b) AIC (c) BIC

Figure 6.8: An all subset analysis for the log-linear model with the cleaned data. All points in the diagram represent a different combination of the subsets and the different shapes of points represent different dummy-families. ND stands for no dummies and is the regression model without any of the dummy families, C indicates that the construction time dummy family is included, L indicates that the location dummy family is included and CL for both the location dummy and construction dummy families included. The optimal solution for the different combinations of dummy-families are marked with a solid line.

In figure (6.8) we observe that the log-linear model has better results for more regressors than the linear model (see figure (6.3)). We now analyse the best models from the different methods using a

cross validation approach (The 10 best models according to the different metrics can be seen in figure (9.5) and (9.6) in the Appendix).

6.2.2 Cross validation analysis of log-linear model

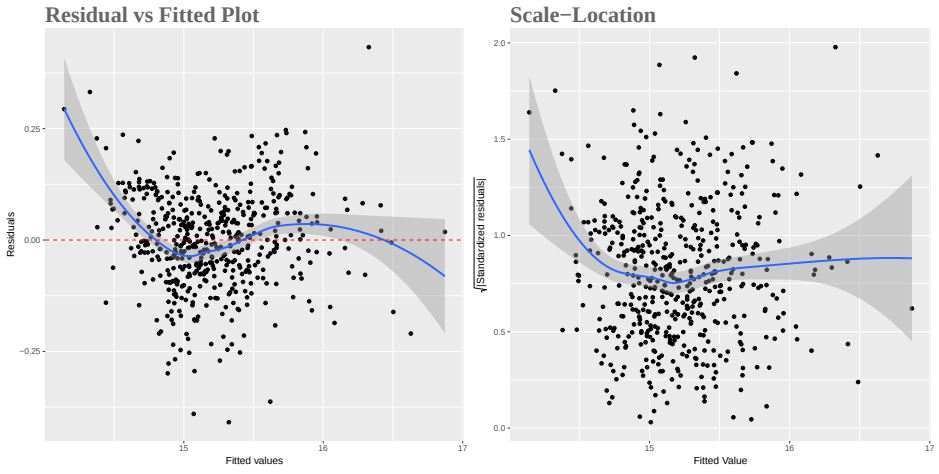
As for the linear model we now perform a cross validation on the 3 best models from the all subsets analysis according to the adjusted R^2 condition, the AIC condition and the BIC condition. We then calculate an R^2 value of the cross validated models and the result can be seen in table (6.3). The performances of all models are very similar but we choose model 6 for further analysis because this model has the best R^2 value.

	SS.Mean	AS.Mean	R2	RA
1	0.0154	0.0965	0.9077	0.7043
2	0.0154	0.0964	0.9077	0.7045
3	0.0154	0.0959	0.9081	0.7061
4	0.0154	0.0960	0.9080	0.7058
5	0.0154	0.0965	0.9077	0.7043
6	0.0154	0.0959	0.9081	0.7061
7	0.0157	0.0967	0.9061	0.7035
8	0.0154	0.0960	0.9080	0.7058
9	0.0158	0.0972	0.9057	0.7021

Table 6.3: Cross validation performed on 9 different models where the first 3 models are the best models according to the adjusted R^2 measure from the all subset regression, model 4 to 6 are the three best models according to the AIC criterion and 7 to 9 is the three best according to the BIC criterion. The cross validation is performed with set.seed() in R which creates the same 10-fold cross validation split for all the models.

6.2.3 A study of the best log-linear model

We now analyse the residuals, the VIF-values and the influential points of the best log-linear model.

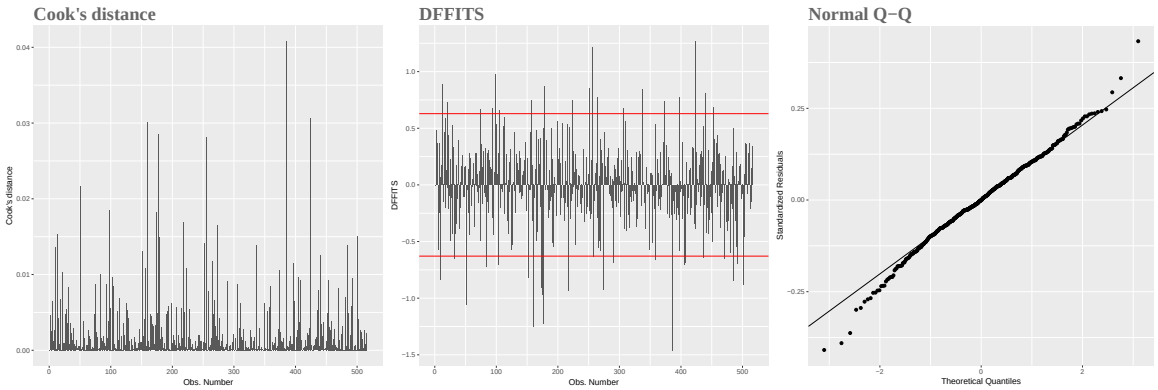


(a) Residuals (b) Norm Residuals

Figure 6.9: The residuals equation (6.3). The model is created with the cleaned data set not containing any missing data.

In figure (6.9) we see that the addition of some square terms and the cluster variables gave smother residuals which is preferable. If we study the QQNorm plot in (6.10) we see that the errors follow

a normal distribution quite well and the transformation in equation (2.16) can therefore be used to estimate an unbiased estimation of the transformed dependent variable. If one analyse the Cook's D plot in the same figure one sees that there exists some data points that affect the model quite heavily. We should therefore consider removing these data points or using a robust regression method to reduce the effect off these data points. A DFBETAS plot is not created as it would need to be created for all regressors in the model.



(a) COOKS D (b) DFFITS (c) QQNorm

Figure 6.10: Cook's D, DFFITS and QQNorm plot for equation (6.3). The model is created with the cleaned data set which does not contain any missing data.

We also investigate possible multicollinearity in the best log-linear model by analysing the VIF-values for the model. We see in table (9.1) in the Appendix that the VIF-values are quite high for most of the continuous variables in the model. We shall therefore test alternative regression models that deal with high multicollinearity (ridge regression, lasso regression and PCR) to see if these models produce better results than the OLS. However one should remember that this model was selected from the all possible regression analysis therefore the multicollinearity should be less of a problem for the prediction.

6.2.4 Models for continued analysis

Based on the analysis above we will continue with 2 models, one log-linear model and one regular model. The focus will be put on the log-linear model as that model has more favourable properties (better fitted values vs residual plot and higher R^2 values in the cross validation analysis) than the linear model. The best models are shown below in equation (6.2) and (6.3):

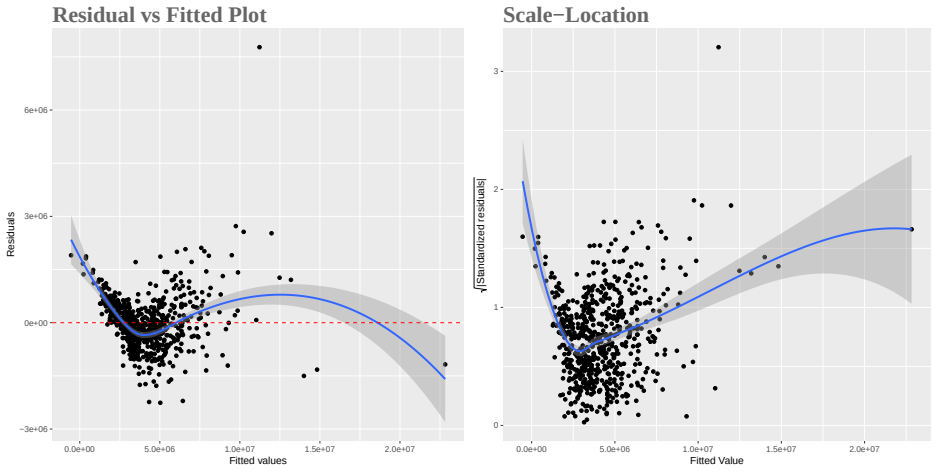
$$
\begin{aligned}
\mathsf{soldPrice}_t = \beta_0 &+ \beta_1 \mathsf{rent}_t \\
&+ \beta_2 \mathsf{livingArea}_t \\
&+ \beta_3 \mathsf{livingArea}_t^2 \\
&+ \beta_4 \mathsf{floor}_t \\
&+ \beta_5 \mathsf{floor}_t^2 \\
&+ \beta_6 \mathsf{location.distance.water}_t \\
&+ \beta_7 \mathsf{location.distance.ocean}_t \\
&+ \beta_8 \mathsf{funkis.CT.dummy}_t \\
&+ \beta_9 \mathsf{folkhem.CT.dummy}_t \\
&+ \beta_{10} \mathsf{miljonprogram.CT.dummy}_t \\
&+ \beta_{11} \mathsf{osubventionerat.CT.dummy}_t \\
&+ \beta_{12} \mathsf{modern.CT.dummy}_t \\
&+ \beta_{13} \mathsf{nyproduktion.CT.dummy}_t \\
&+ \beta_{14} \mathsf{missing.CT.dummy}_t + \\
&+ \sum_{cluster} \beta_i \mathsf{clusterdummy}_{it} + \epsilon
\end{aligned}
\tag{6.2}
$$

$$
\begin{aligned}
log(\mathsf{soldPrice}_t) = \beta_0 &+ \beta_1 \mathsf{rent}_t \\
&+ \beta_2 \mathsf{rent}_t^2 \\
&+ \beta_3 \mathsf{livingArea}_t \\
&+ \beta_4 \mathsf{livingArea}_t^2 \\
&+ \beta_5 \mathsf{rooms}_t \\
&+ \beta_6 \mathsf{rooms}_t^2 \\
&+ \beta_7 \mathsf{floor}_t \\
&+ \beta_8 \mathsf{floor}_t^2 \\
&+ \beta_9 \mathsf{location.distance.ocean}_t \\
&+ \beta_{10} \mathsf{funkis.CT.dummy}_t \\
&+ \beta_{11} \mathsf{folkhem.CT.dummy}_t \\
&+ \beta_{12} \mathsf{miljonprogram.CT.dummy}_t \\
&+ \beta_{13} \mathsf{osubventionerat.CT.dummy}_t \\
&+ \beta_{14} \mathsf{modern.CT.dummy}_t \\
&+ \beta_{15} \mathsf{nyproduktion.CT.dummy}_t \\
&+ \beta_{16} \mathsf{missing.CT.dummy}_t + \\
&+ \sum_{cluster} \beta_i \mathsf{clusterdummy}_{it} + \epsilon
\end{aligned}
\tag{6.3}
$$

6.2.5 Handling of missing data

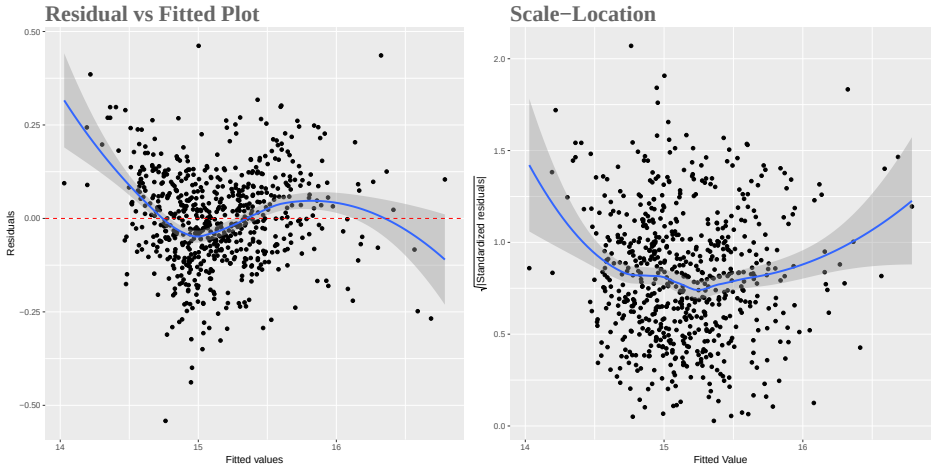
A model that contains missing data can lead to a biased model, so here we will investigate the method of adding a dummy variable for all of the variables that have a problem of missing data, the dummy variable has the value 1 if data is missing and 0 otherwise. We then set all the missing data points to 0 and let the impact on the model be through the dummy variable. The benefit of this method is that less data is discarded in the cleaning process.

After analysing figure (4.2) we make the decision to evaluate a dummy variable fix for location.distance.ocean and floor (constructionYear already has a missing data dummy in the dummy variable family). The reason that we do not create missing dummy variables for more variables is that those variables have so few data points missing. Creating a missing dummy variable could then lead to a column in the independent variable matrix $\mathbf{X}$ which just consists of 0s, which would lead to an non invertible matrix.



(a) Residuals (b) Norm Residuals

Figure 6.11: The residuals from equation (6.2) with dummyFix variables added for location.distance.ocean and floor. The full data set now only contained 13 missing data points that were excluded from the model.



(a) Residuals (b) Norm Residuals

Figure 6.12: The residuals equation (6.3). with dummyFix variables added for location.distance.ocean and floor. The full data set now only contained 13 missing data points that were excluded from the model.

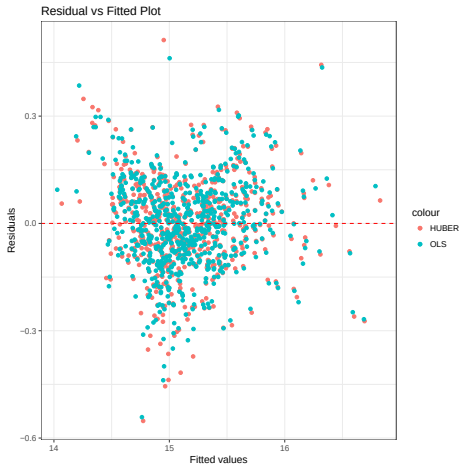
We see that the residuals of the models with the dummyFix is similar to the models with the cleaned data (compare figure (6.11) with figure (6.5) and figure (6.12) with figure (6.9) to see the similarities for both the regular and log-linear model). We will therefore continue the analysis with the models with the dummy fix for the missing data.

6.3 Other regression methods

We now examine the results of using robust regression methods to deal with outliers and high multicollinearity. We will conduct some analysis of some models that are described in the theory section and see if these models improve the linear valuation model.

6.3.1 Huber regression

The method rlm in R is used to perform Huber regression on the model from equation (6.3). The method finds the solution by using an iterative reweighing process.



(a) log-linear model

Figure 6.13: The residuals for the Huber regression and the OLS regression methods using the model from equation (6.3).

A Huber regression has different weighting function than the OLS regression and the Huber weighting takes outlier into less consideration than the OLS method. The Huber method will most likely lead to some more extreme outliers than the OLS method. (OLS gives the same weight to the outliers which give them a better fit but hurts the robustness of the model). This can clearly be seen in figure (6.13).

A cross validation analysis is also performed for the OLS and Huber regression and the results can be seen in table (6.4). We see that the cross validation results are very similar for the OLS and Huber regression, but the model using the OLS regression method performs marginally better for the squared error measures and the model using the Huber regression method performs marginally better on the absolute deviation measures.

	SS.Mean	AS.Mean	R2	RA
OLS	0.0197	0.1091	0.8882	0.6769
HUBER	0.0198	0.1083	0.8875	0.6794

Table 6.4: A cross validation for the Huber regression and the OLS regression methods using the model from equation (6.3).

The results for the OLS and Huber regression are very similar for this time period even that the influential points analysis done earlier showed that there exist many influential points in the data. One could therefore argue that the theoretical properties (less influence of outlier points and therefore a more robust model) of the Huber regression still makes it a interesting parameter estimation method.

We will therefore model and test indexes using a Huber regression method.

6.3.2 Ridge regression

In table (6.3) we see that the best log-linear model suffers from a problem with high multicollinearity. As mentioned in the theory section the ridge regression method (see equation (2.25)) is a potential solution for models with high multicollinearity so this method will now be examined further. We start by performing a ridge regression with the standardized variables and the result from the glmnet() function from the glmnet package can be seen in figure (6.14).

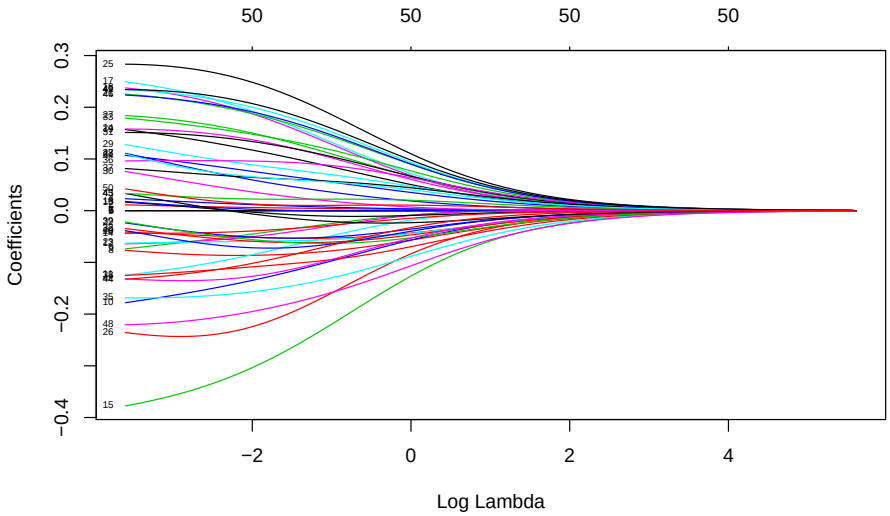


Figure 6.14: The standardized coefficients for the ridge plot for the model from equation (6.3). The numbers over the plot are the number of coefficients that are still left in the model for the corresponding value of the logarithm of the penalty term $\log(\lambda)$.

We now perform a cross validation do decide the best value for λ for the ridge regression model. We use the function cv.glmnet() in the glmnet package to do this cross validation.

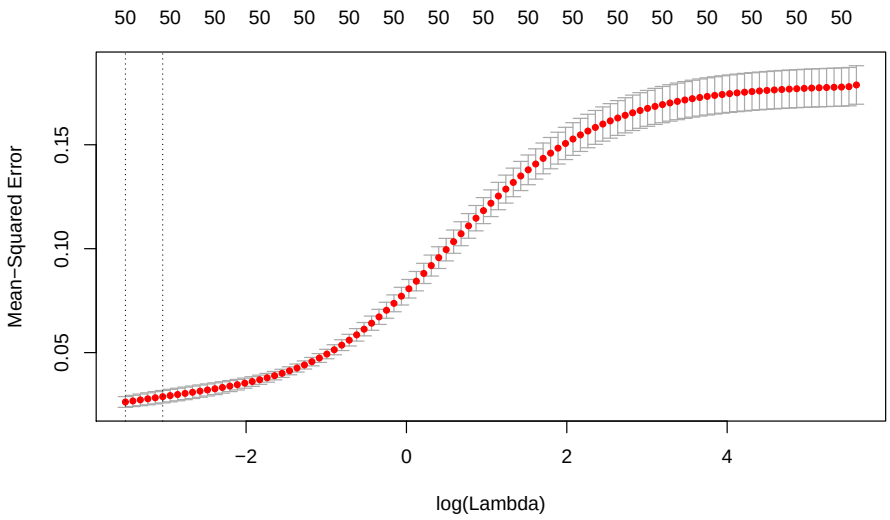


Figure 6.15: The cross validation result for the mean square error of the ridge plot for the model from equation (6.3). The numbers over the plot are the number of coefficients that are still left in the model for the corresponding value of the logarithm of the penalty term $\log(\lambda)$.

From the cross validation of the ridge model in figure (6.15) we see that the model takes on small mean squared error values for small values of λ. This is a sign that we do not need the ridge regression as a λ that goes towards 0 is equivalent to the OLS method. This result is further strengthened by the shape of the coefficient in figure (6.14) which show that all variables grow coherently. This is a sign that the ridge regression does not provide a large benefit over the OLS regression. A ridge regression would therefore just contribute to a more complicated model which is not favourable. We will therefore not use the ridge regression method to model any RPPIs.

6.3.3 Lasso regression

We now investigate the lasso regression method which can be seen in equation (2.29). The penalty term in the lasso regression is strictly positive so this method will reduce the model and remove coefficients when the penalty term is large. This can be seen in figure (6.16) where the model only contains few regressors for large values of λ.

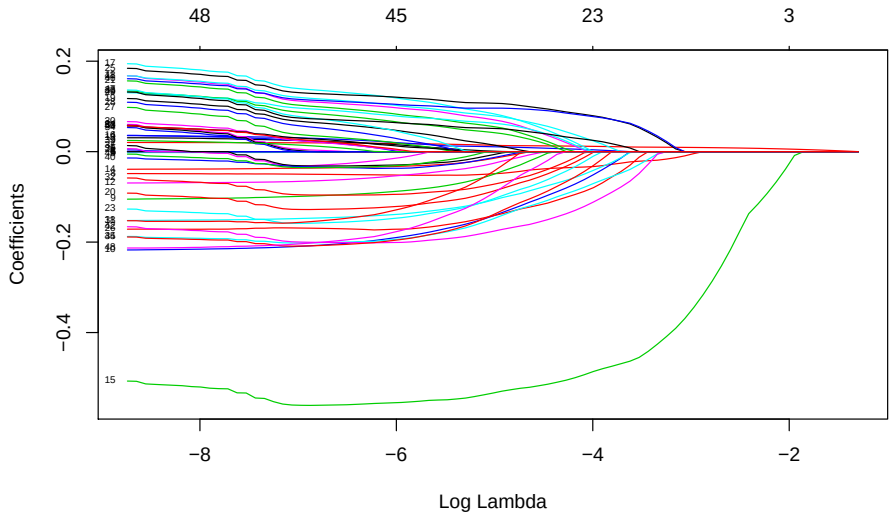


Figure 6.16: The standardized coefficients for the lasso plot for the model from equation (6.3). The numbers over the plot are the number of coefficients that are still left in the model for the corresponding value of the logarithm of the penalty term $\log(\lambda)$.

We now perform a cross validation to decide the best value for λ for the lasso regression model. We use the function cv.glmnet() in the glmnet package to do this cross validation.

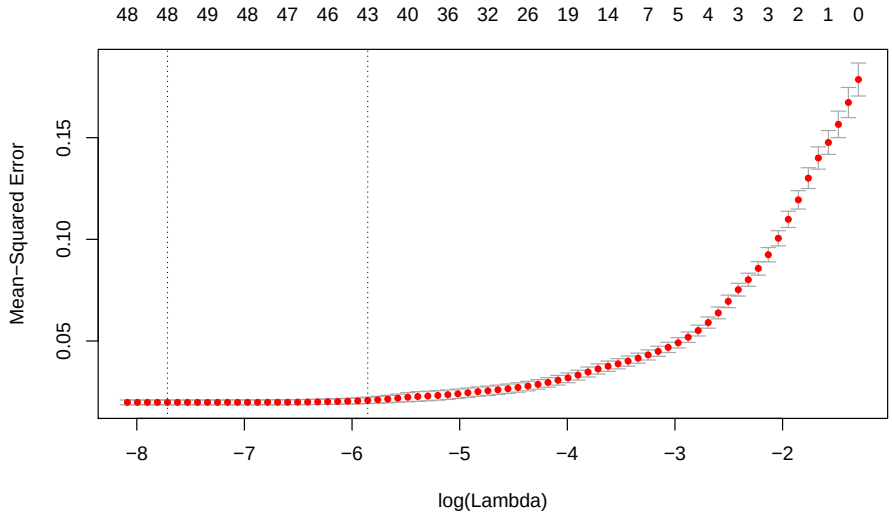
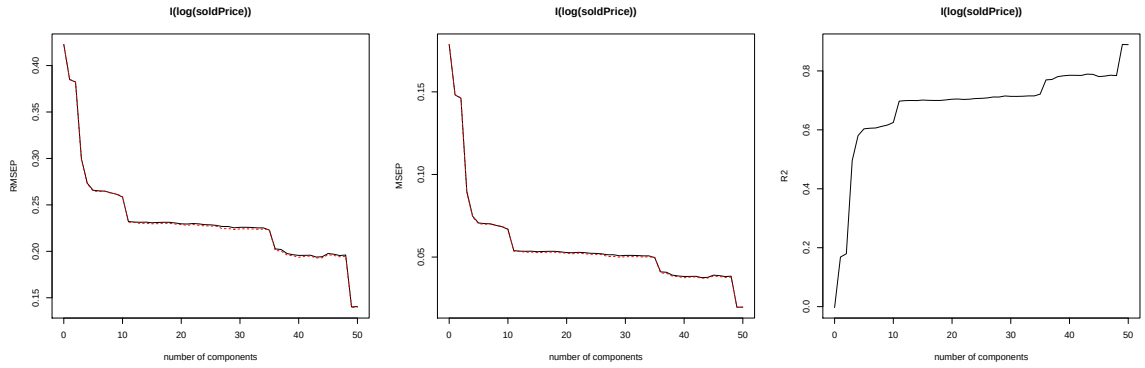


Figure 6.17: The cross validation result for the mean square error for the lasso plot for the model from equation (6.3). The numbers over the plot are the number of coefficients that are still left in the model for the corresponding value of the logarithm of the penalty term $\log(\lambda)$.

From figure (6.17) we see that the lasso model takes the lowest squared errors for small values of lambda which is an indication that the penalty in the lasso regression does not improve the results of the model significantly. It would therefore not be beneficial to use the lasso regression model in construction of the index as it would only introduce more complexity to the model.

6.3.4 PCR

We perform a PCR with the pcr method from the pls package in R. We perform a 10-fold cross validation on the PCR with all the different number of PCA components. The PCR is performed for the model from equation (6.3) and the result is shown in figure (6.18).



(a) root mean squared error (b) mean squared error (c) R^2

Figure 6.18: The result for the PCR for the model from equation (6.3). The result is for a 10 fold cross validation and (a) shows the root mean squared error for the PCR, (b) shows the mean squared error for the PCR and (c) the R^2 for the PCR.

We see that all the performance metrics in figure (6.18) improve for an increasing number of PCA-components. According to [15] this is a sign that PCR should not be used. If the PCR produces a low cross validation error with few PCs the PCR is the best method to use, but if PCR produces the smallest cross validation results for a number of PCs that is close to the actual number of regressors (as in our case) the OLS method is superior to PCR.

We have now tested ridge regression, lasso regression and PCR and these methods did not improve the results from the OLS method. This is a sign that the multicollinearity that exist in the model does not create problems (because if it did the methods reducing the multicollinearity problem would have produced better results than the OLS method). We will therefore not use these methods in the index creation in the coming sections.

6.4 Modelling of the index

In the previous sections we found that the log-linear model using OLS regression and the log-linear model using Huber regression performed well after performing analysis on the model for December 2016. We now model and analyse the indexes according to the different methods discussed in the method segment. We will model the RPPIs for Stockholm during the period for January 2013 to December 2016

6.4.1 Average approach

We here model the RPPIs using the average price approach. We model 2 versions of it, one with all the data and one with only the data that is used in the characteristic and hedonic models in the later subsections. The result is shown in figure (6.19) and we can see that the average approach produce an index that is very volatile.

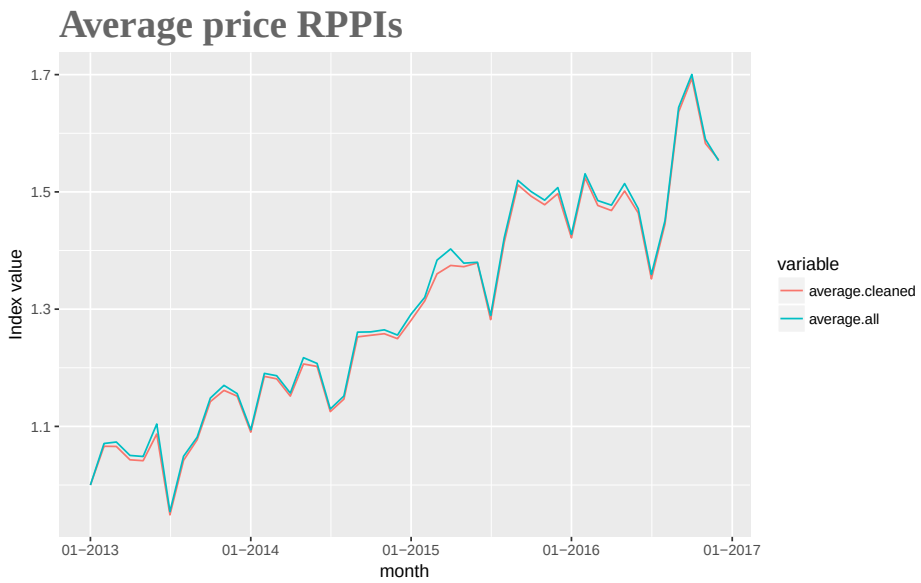


Figure 6.19: The RPPIs modeled using the average price approach for all the available data (average.all) and for the data used in the characteristic approach methods (average.cleaned).

The volatility comes from the fact that the average price RPPI can not compensate for the differences in the characteristics (See figure (6.30) and (6.31))of the apartments sold in the period. This is the main purpose of the characteristic/hedonic approaches that will be tested next.

6.4.2 Characteristic/hedonic approach

We now use the OLS and HUBER regression methods to model RPPIs using a Fisher type characteristic approach, which is the same as the hedonic double imputation method for linear models (see equation (5.7)). The best log-linear model from equation (6.3) will be used in all periods and the cluster dummy variables will be chosen so that there exist at least 5 entries in every cluster for all time periods of the index, see model 1 from the method section.

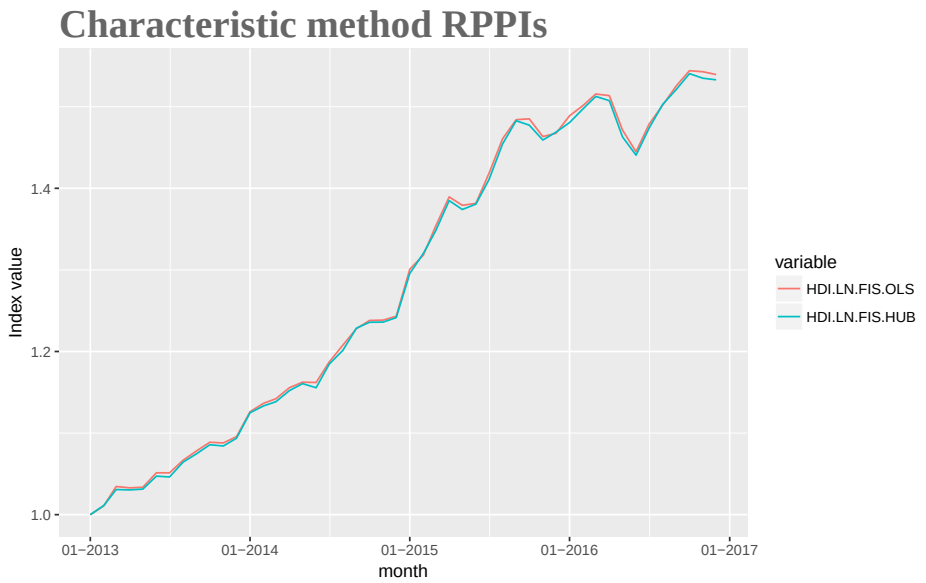
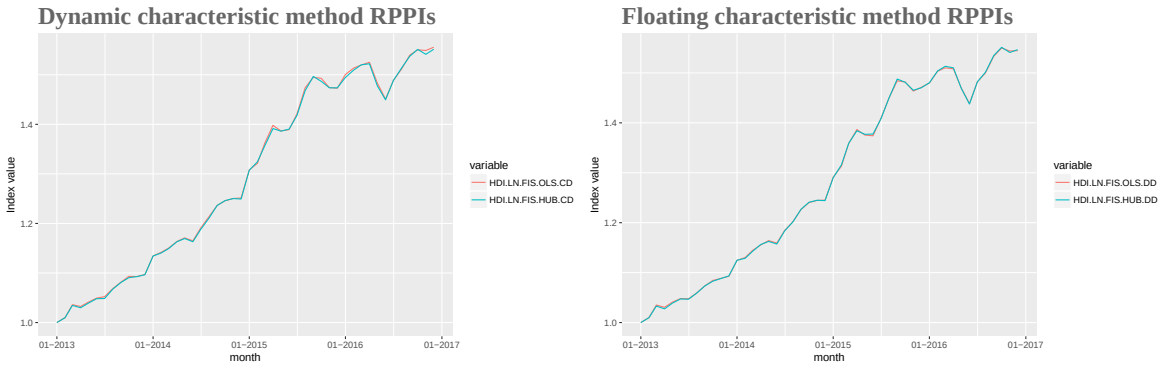


Figure 6.20: The RPPIs modeled using the characteristic/hedonic approach for (model 1) of equation (6.3) using OLS (HDI.LN.FIS.OLS) and Huber regression (HDI.LN.FIS.HUB).

One can see that the Characteristic/hedonic imputation approach (in figure (6.20)) produces less volatile RPPI than the average approach (in figure (6.19)). Figure (6.20) also show that there is not a big difference between the OLS and HUBER regression methods when it comes to the modeled index.

6.4.3 Dynamic Characteristic approach

Model 2 and model 3 for the characteristic/double imputation approach will now be examined. Model 2 is named with the ending CD which stands for cluster dynamic and model 3 with the ending DD which stands for double dynamic. The RPPIs is modeled using both OLS and HUBER for both model 2 and 3 which can be seen in figure (6.21)



(a) (model 2)

(b) (model 3)

Figure 6.21: The RPPIs modeled using the characteristic/hedonic approach for equation (6.3) using OLS (HDI.LN.FIS.OLS) and Huber regression (HDI.LN.FIS.HUB). (a) shows the index for (model 2) from the method section where the CD stands for cluster dynamic. (b) shows the index for (model 3) from the method section where the DD stands for Double Dynamic (both the clusters and the way the index is calculated)

We now plot the differences for the different methods of the characteristic approach in a graph where

the first static characteristic model (model 1) is the base value. We create the plot for both the OLS and HUBER regression models.

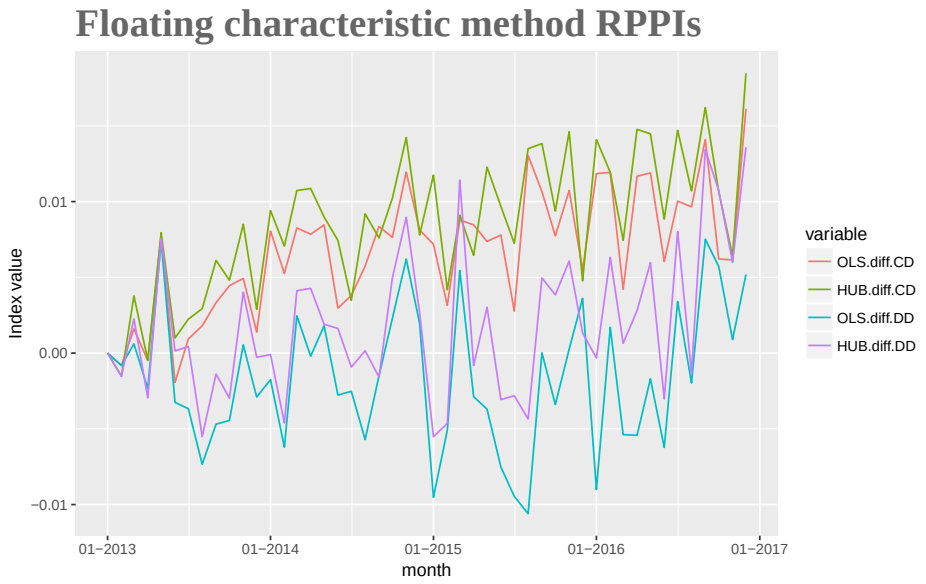


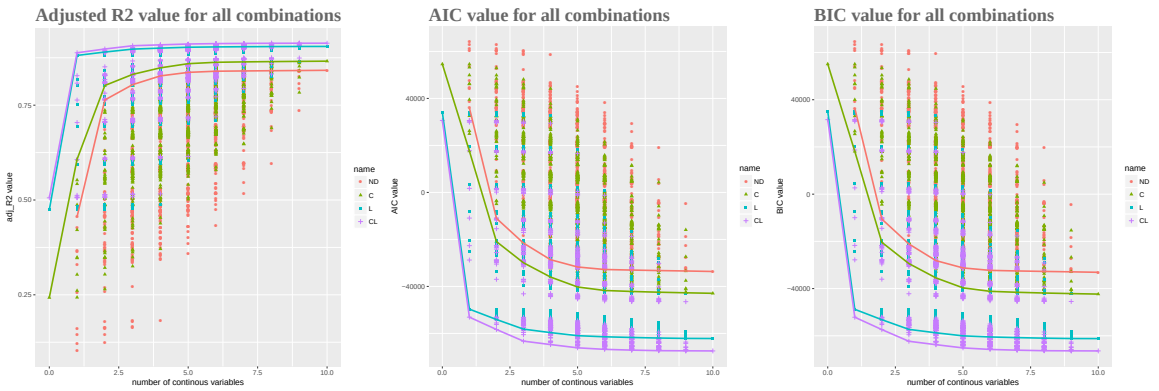
Figure 6.22: The differences in model 1-3 for the characteristic/double imputation approach. Model 1 is the base case and the differences for model 2 (CD) and model 3 (DD) is plotted for the OLS and HUBER regression methods

We see in figure (6.22) that the difference for the different index models is not substantial as the difference is between -0.01 and 0.015 which corresponds to a difference of a few percent. But the models will be examined further in the index validation section.

6.4.4 Creating the time-dummy index

Here we investigate methods to find a "best" version of the time dummy index. We perform an all possible subsets regression in the same way as we did for the December 2016 data and choose the best log-linear time dummy model from that analysis. As we found in the previous analysis we want to use the model with missing variable dummies. The all possible subset regression will be done with missing variable dummies for floor and location.distance.ocean. The dummy fix variable floor.missingDummy is included if the set of variables contains floor or floor.2 and location.distance.ocean.missingDummy is included if the variable string contains location.distance.ocean.

One thing to keep in mind when this analysis is performed is to remove the first cluster dummy variable as well to avoid perfect collinearity between the cluster dummy variables. This was not a problem for the double hedonic imputation indexes as there was always some object that belonged to a cluster that had less than 5 data points and those cluster points became the base case (but one should look out for perfect collinearity in those cases as well).



(a) Adj R2 (b) AIC (c) BIC

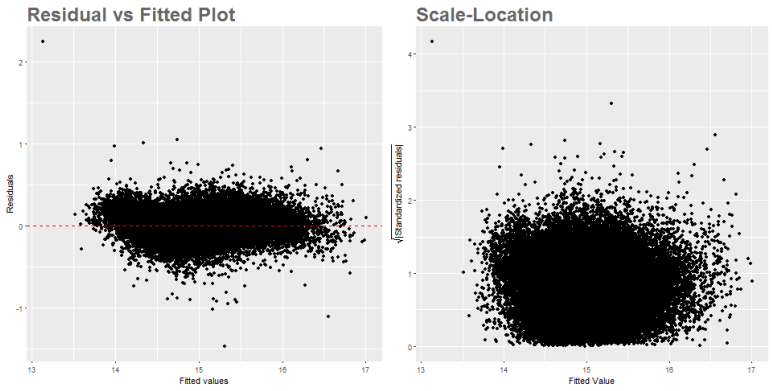
Figure 6.23: An all subset analysis for the log-linear time model with the cleaned data. All points in the diagram represent a different combination of the subsets and the different shapes of points represent different dummy-families. ND stands for no dummies and is the regression model without any of the dummy families, C indicates that the construction time dummy family is included, L indicates that the location dummy family is included and CL for both the location dummy and construction dummy families included. The optimal solution for the different combinations of dummy-families are marked with a solid line.

We can see the result from the all possible subset regression model in figure (6.23) and figure (9.7) and (9.8) in the Appendix. As before we choose the three best models according to the three different matrices and perform a cross validation analysis on these models. We see that the different evaluation methods (adj.r2 ssr mm) give the same best models so we test these 10 best models using a 10 fold cross validation.

	SS.Mean	AS.Mean	R2	RA
1	0.0201	0.1013	0.9010	0.7177
2	0.0202	0.1013	0.9010	0.7176
3	0.0179	0.1010	0.9122	0.7184
4	0.0179	0.1011	0.9122	0.7183
5	0.0203	0.1016	0.9002	0.7169
6	0.0178	0.1013	0.9126	0.7176
7	0.0203	0.1016	0.9001	0.7168
8	0.0202	0.1015	0.9006	0.7171
9	0.0178	0.1014	0.9125	0.7175
10	0.0202	0.1015	0.9005	0.7170

Table 6.5: A cross validation analysis done on the 10 "best" models from the all subset analysis from (6.23) and figure (9.7) and (9.8) in the Appendix.

From table (6.5) we see that model 6 has the largest R^2 value and smallest SS.Mean value and is therefore the best model that we are going to continue our analysis with. We now analyse the error plot for this model and the results can be seen in figure (6.24). We see that the residuals vs fitted values plot has an even shape so we now want to check if there exists many outlier points in the dataset. The previous method with plotting Cook's D and DFFITS for all points is not applicable now because the dataset is too large. However we can see that the residual plots contain some outlier values and we therefore conclude that it would be interesting to use the robust HUBER regression method as a complement to the OLS regression method in modeling the index using the time dummy method.



(a) Residuals (b) Standardized residuals

Figure 6.24: The residuals (a) and the standardized residuals (b) for the "best" time dummy variable model that can be seen in equation (6.4)

The potential multicollinearity of the time dummy variables is the most interesting issue in this model as high multicollinearity in these variables would lead to an unstable model and therefore an unrobust index. Table (6.6) show the VIF values for the time dummy variables and we can see that the maximum VIF value is 3.04 which is below the critical Value. We therefore will not investigatet the multicollinearity fixing models (PCR, Lasso and Ridge).

	VIF		VIF
dummy.201302	2.09	dummy.201502	1.9
dummy.201303	2.28	dummy.201503	2.73
dummy.201304	2.15	dummy.201504	2.61
dummy.201305	2.18	dummy.201505	2.91
dummy.201306	1.98	dummy.201506	2.53
dummy.201307	1.46	dummy.201507	1.58
dummy.201308	2.03	dummy.201508	2.26
dummy.201309	2.15	dummy.201509	2.56
dummy.201310	2.34	dummy.201510	2.64
dummy.201311	2.23	dummy.201511	2.55
dummy.201312	1.61	dummy.201512	1.79
dummy.201401	2.04	dummy.201601	2.19
dummy.201402	2.18	dummy.201602	2.6
dummy.201403	2.10	dummy.201603	2.62
dummy.201404	2.31	dummy.201604	3.04
dummy.201405	2.56	dummy.201605	2.71
dummy.201406	2.15	dummy.201606	2.24
dummy.201407	1.49	dummy.201607	1.61
dummy.201408	2.26	dummy.201608	2.32
dummy.201409	2.58	dummy.201609	2.71
dummy.201410	2.69	dummy.201610	2.74
dummy.201411	2.40	dummy.201611	2.6
dummy.201412	1.83	dummy.201612	1.72
dummy.201501	2.10		

Table 6.6: The VIF-values for the time dummy variables from equation (6.4)

$$
\begin{aligned}
log(\text{soldPrice}_t) = \beta_0 &+ \beta_1 \text{rent}_t \\
&+ \beta_2 \text{livingArea}_t \\
&+ \beta_3 \text{livingArea}_t^2 \\
&+ \beta_4 \text{rooms}_t \\
&+ \beta_5 \text{floor}_t \\
&+ \beta_6 \text{floor}_t^2 \\
&+ \beta_7 \text{location.distance.water}_t \\
&+ \beta_8 \text{location.distance.ocean}_t \\
&+ \beta_9 \text{funkis.CT.dummy}_t \\
&+ \beta_{10} \text{folkhem.CT.dummy}_t \\
&+ \beta_{11} \text{miljonprogram.CT.dummy}_t \\
&+ \beta_{12} \text{osubventionerat.CT.dummy}_t \\
&+ \beta_{13} \text{modern.CT.dummy}_t \\
&+ \beta_{14} \text{nyproduktion.CT.dummy}_t \\
&+ \beta_{15} \text{missing.CT.dummy}_t \\
&+ \beta_{16} \text{floor.missingDummy}_t \\
&+ \beta_{17} \text{location.distance.ocean.missingDummy}_t \\
&+ \sum_{\text{cluster}} \beta_i \text{clusterdummy}_{it} \\
&+ \sum_{\text{time periods}} \gamma_j \text{timeDummy}_{jt} + \epsilon
\end{aligned}
\tag{6.4}
$$

The best time dummy model from the analysis from above gives us the following index that can be seen in figure (6.25).

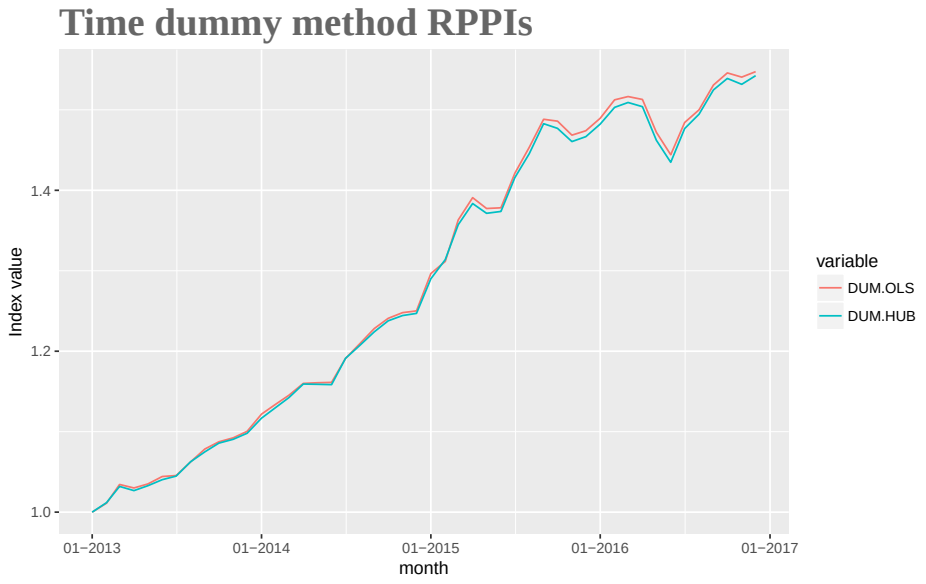


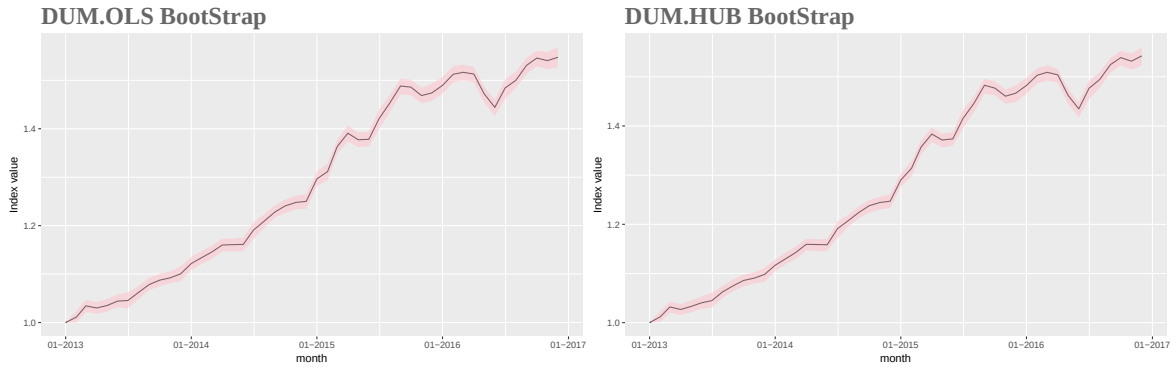
Figure 6.25: The time dummy Index modeled using equation (6.4) using both an OLS and HUBER regression approach

6.5 Validation of the index

We shall now evaluate different methods to validate the different index modeling methods. We will begin by performing the bootstrap validation method that was mentioned in the methodology section. The coefficients for the continuous variables will then be plotted to analyse the differences on the index that is introduced by using OLS or HUBER regression in the construction of the index. Mean value of the characteristics will then be investigated to see if the time dummy method and double imputation method give similar results.

6.5.1 Bootstrap validation

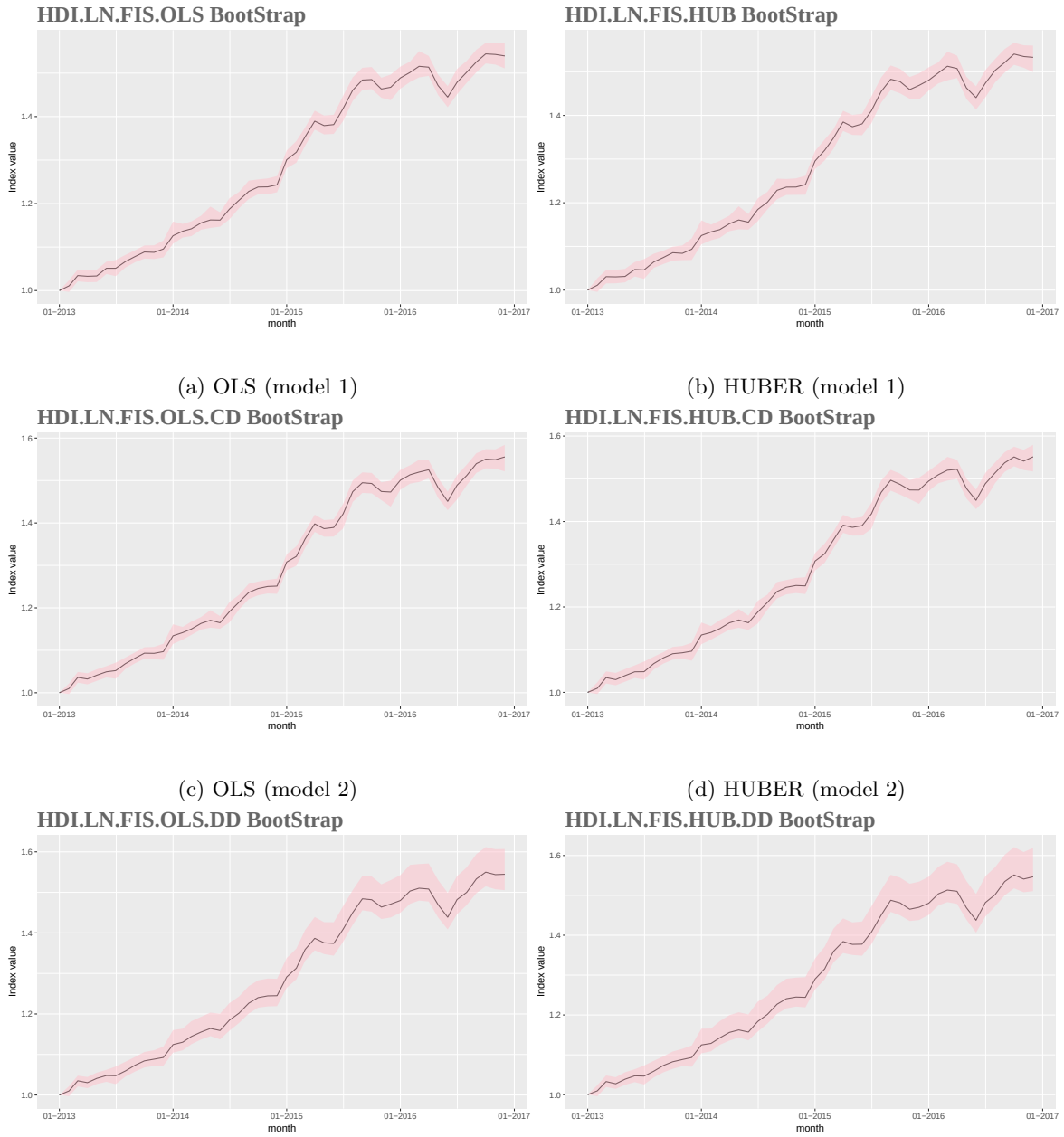
We here implement a validation approach using Bootstrapping described in the methods section. We start by plotting the modeled index with a empirical 95% confidence interval calculated by creating 1000 bootstrap samples and order the index values for all the time points and taking the 0.025 and 0.975 percentile. The reason a larger boostrap sample was not used is the computational complexity of calculating the index and the complexity of creating a bootstrap sample with a size of $n > 56000$.



(a) Time dummy OLS

(b) Time dummy HUBER

Figure 6.26: The RPPIs with empirical Bootstrap confidence intervals modeled using the time dummy approach for equation (6.4) (a) shows the index for OLS time dummy regression method (b) shows the index for HUBER time dummy regression method



(a) OLS (model 1)

(b) HUBER (model 1)

(c) OLS (model 2)

(d) HUBER (model 2)

(e) OLS (model 3)

(f) HUBER (model 3)

Figure 6.27: The RPPIs with empirical Bootstrap confidence intervals modeled using the characteristic/hedonic approach for equation (6.3). (a) shows the index for OLS regression and (model 1) from the method section. (b) shows the index for HUBER regression and (model 1) from the method section. (c) shows the index for OLS regression and (model 2) from the method section. (d) shows the index for HUBER regression and (model 2) from the method section. (e) shows the index for OLS regression and (model 3) from the method section. (f) shows the index for HUBER regression and (model 3) from the method section.

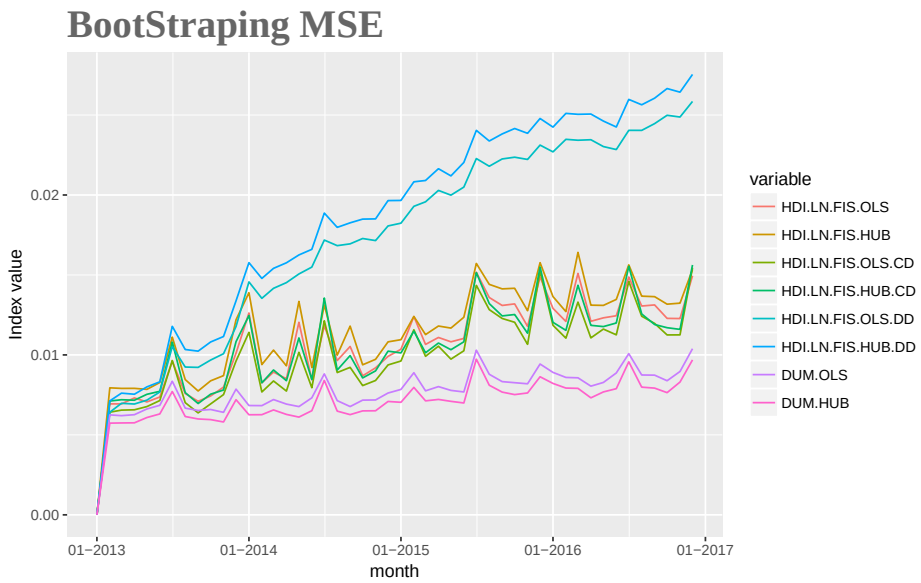


Figure 6.28: Approximation of MSE using equation (5.12) with 1000 re samples for the different RPPIs methods. HDI stands for hedonic double imputation and DUM for time dummy method. CD stands for Cluster Dynamic and is model 2 and DD stands for double dynamic and is model 3.

In figure (6.28) we can see that model 3 produces unrobust results as the MSE increases with time. This is a sign of error propagation and is not favourable. This can also bee seen in figure (6.27) as model 3 has the largest confidence interval. We also see that the time dummy HUBER model has a lower MSE than the OLS time dummy model and that the best hedonic double imputation model is the HDI.LN.FIS.HUB.CD model and we will therefore plot these two versions of the RPPI in the same figure with the confidence intervals included to see the similarities and differences between these two models.

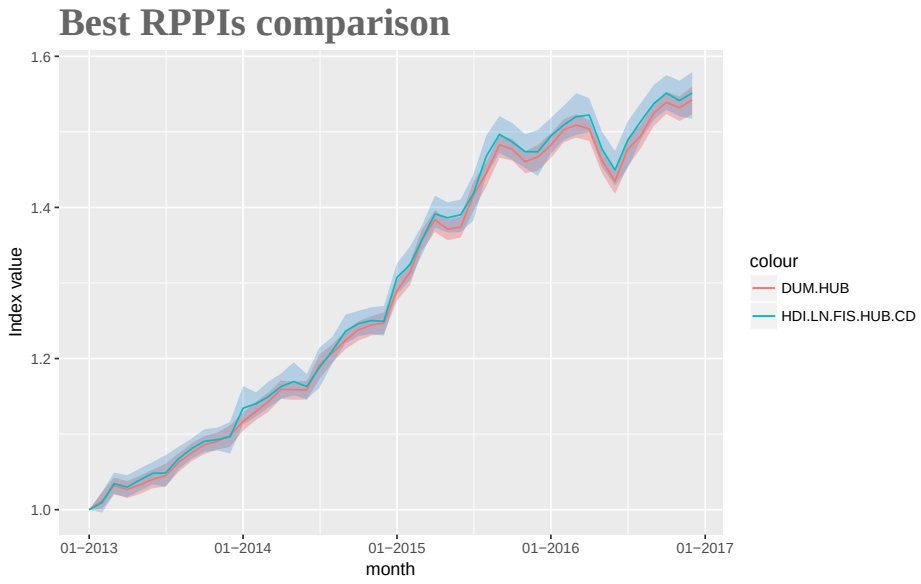


Figure 6.29: The best time dummy model (DUM.HUB, which is the time dummy model with a HUBER regression method described earlier) and the best double imputation model (HDI.LN.FIS.HUB.CD, which is the double imputation model 2 with HUBER regression from above). The indexes are plotted with the empirical 95% confidence intervals marked with a semitransparent region.

In figure (6.29) we see that the time dummy approach and the double hedonic approach produce quite similar index results even if some noticeable differences exist. It should be noticed that the 95% confidence interval for both methods are overlapping meaning that one can not say with 95% confidence level the methods produce different results. Also worth mentioning is the fact that the double hedonic imputation index is higher than the time dummy index for all times. The reason for this can be a change in the underlying characteristics which will be investigated later in the report.

6.5.2 Comparison of the coefficients

In this section we compare the different regression methods used in the modeling of the hedonic indices to see which has the most preferable attributes. We plot the coefficients over the different time periods and for the method to be as robust as possible we want the graphs to be as smooth as possible. This comparison is done for model 1 as using the same model in every time period give more comparable results, the change of included variables in model 2 or 3 could introduce some of the volatility in the coefficients otherwise.

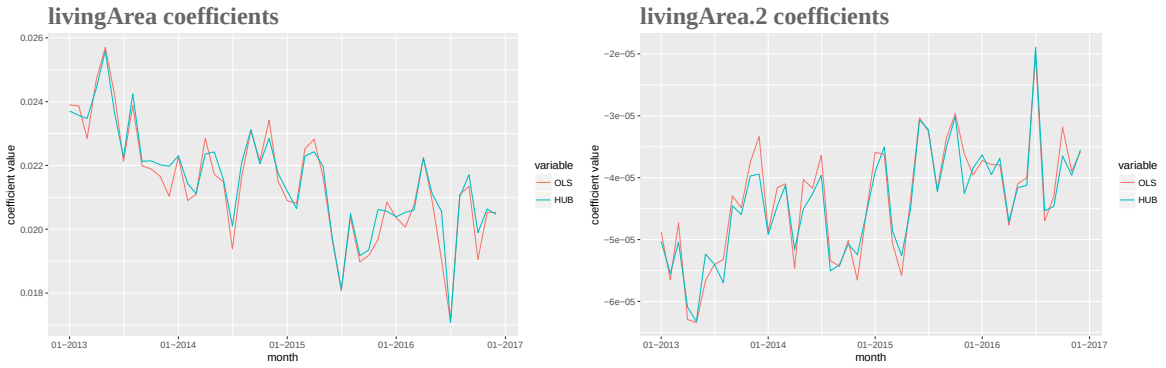
(a) constant

(b) rent

(c) floor

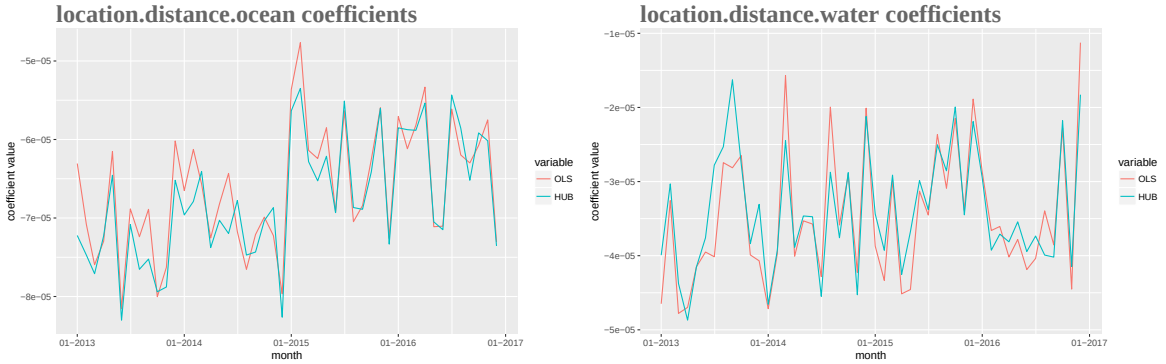
(d) floor.2

Figure 6.30: The continuous coefficients plotted for the hedonic imputation model 1 for the different time periods. The OLS and Huber (HUB) regression methods are plotted. part(1 of 2)



(a) livingArea

(b) livingArea.2



(c) location.distance.ocean

(d) location.distance.water

Figure 6.31: The continuous coefficients plotted for the hedonic imputation model 1 for the different time periods. The OLS and Huber (HUB) regression methods are plotted. part(2 of 2)

We see in figure (6.30) and (6.30) that the Huber regression method performed better than the OLS regression method in regards to robustness but both methods produce quite volatile coefficients and the difference is insignificant. One can also see that the constant variable in the models has a rising trend similar to the calculated RPPIs.

6.5.3 Comparison of the characteristics

In this section we plot the mean value of the characteristics for every period. This is done to examine how appropriate it is to use the time dummy index method. The time dummy method and the double imputation method will produce similar indexes if the underlying characteristics is fairly constant over different time periods. However as mentioned in the literature background a double imputation approach is superior if the underlying characteristics change substantially during the time period.

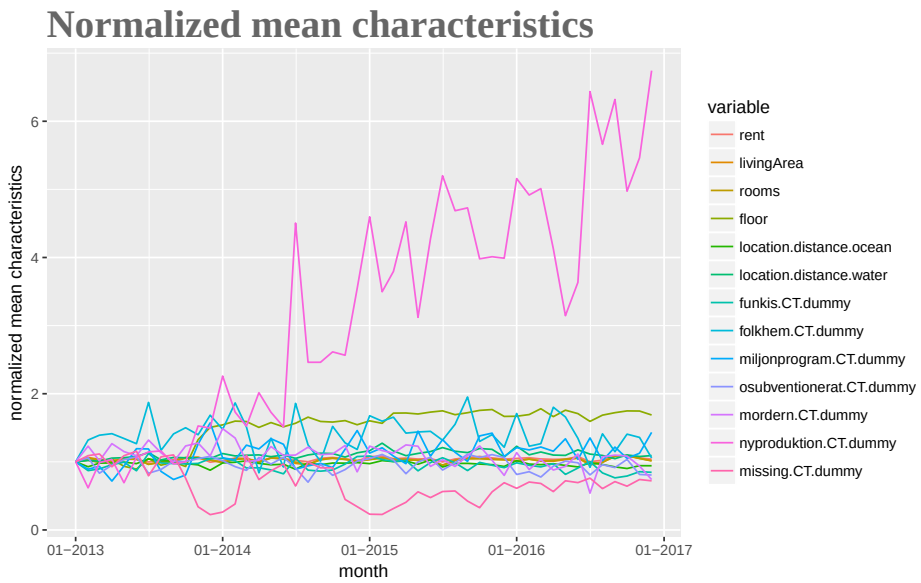


Figure 6.32: Normalized mean characteristics for all the continuous variables and the constructionDummy family

Figure (6.32) contains the outlier nyproduction.CT.dummy so we remove all the construction time dummies and plot the mean values of the characteristics again to get a better look at the other variables.

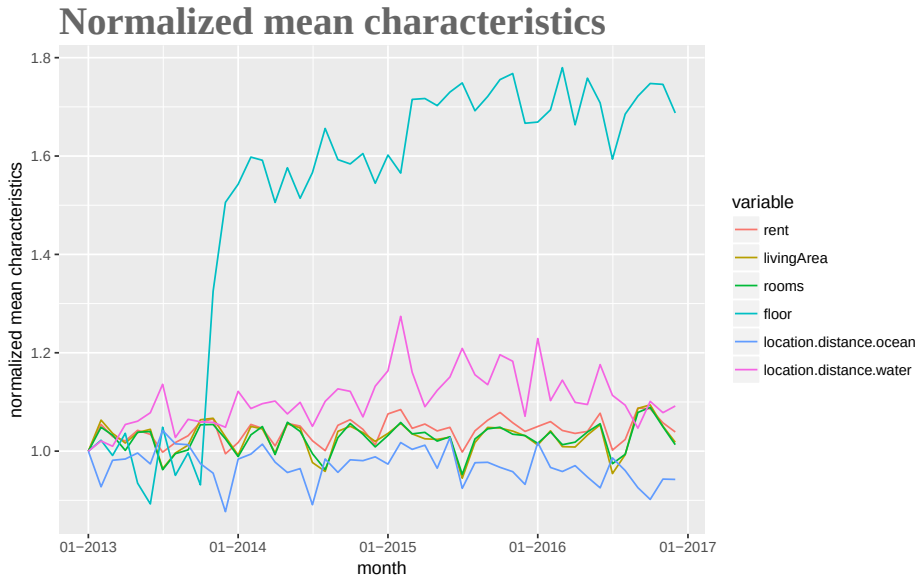


Figure 6.33: Normalized mean characteristics for all the continuous variables

In figure (6.32) and (6.33) we see that the characteristics of the apartments are relatively constant during the different time periods. The two outliers are nyproduktion.CT.dummy and floor.

6.6 Amount of data

In the previous sections we found that the Huber regression time dummy index (DUM.HUB) and the double imputation model 2 index (HDI.LN.FIS.HUB.CD) are the best time dummy and double

imputation approach index modeling methods. We will now test how these index methods perform with a smaller dataset that is simulated from the original Stockholm 2013-2016 dataset. This will be done to examine how the index methodologies perform for smaller datasets and with this method one could compare the smaller dataset result with the full data set result. The smaller data set will not have exactly the same mix of characteristics as the full dataset so the indexes will therefore differ. But most of the difference will come from unrobustness in the index method when the index is modeled from a smaller dataset.

We will analyse data sets which contains 75%, 50%, 25%, 10% and 5% of the original data and where the data is selected randomly and without re-sampling of the same data points. The best models will be decided using the all possible subset regression and choosing the best index according to the AIC criterion. The index is then modeled and comparisons are done with the index modeled with the full dataset. We start by plotting the number of data points in the different time periods for the 5 newly created datasets to see how the data is distributed in the different time periods.

	min	1st quartile	median	3rd quartile	max
5%	18	43	64	75	103
10%	46	97	116	142	210
25%	107	255	285	351	468
50%	216	496	594	735	950
75%	329	714	864	1110	1463

Table 6.7: A five number statistic table over the amount of data from the reduction of the full dataset. The number of data points in each period can be seen in figure (6.34)

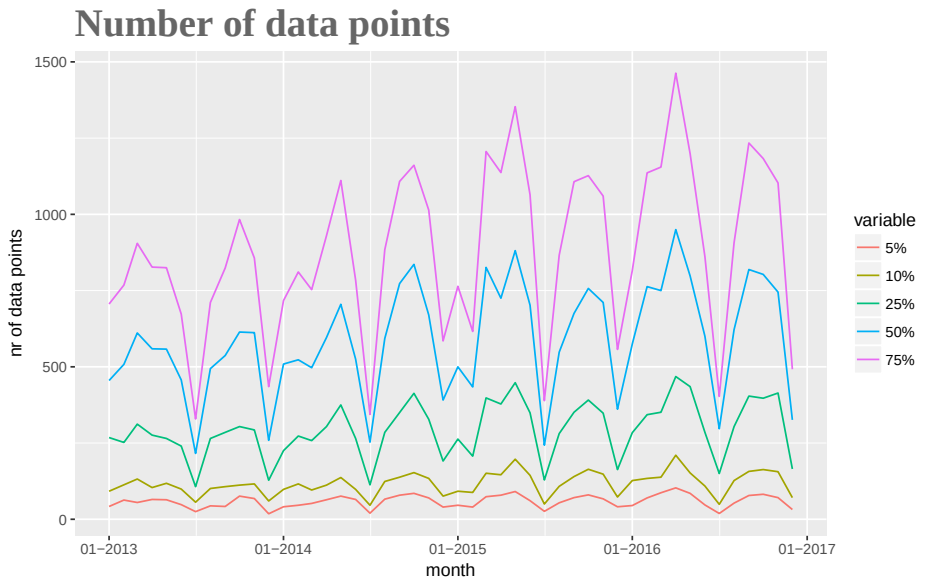


Figure 6.34: The number of data points in each time period for the reduction of the main data source. The data is also described in a five number analysis in table (6.7) to get an exact numerical range of the number of data points in the different time periods.

An all subset regression analysis is now performed on each of the stratified datasets to see which model to choose for the different datasets. The result can be seen in figure (9.9), (9.10) and (9.11) in the Appendix. One can clearly see that the valuation model for the hedonic double imputation index with only 5% of the data is not optimal as it only contains 4 variables. This can also be seen in figure (6.35) which show that the HDI.5%.diff differs substantially from the index modeled with all the data.

The index should deviate from the index modeled with the full data amount as the underlying data represent a different subset of apartments but the difference in the time dummy index in figure (6.36) is much smaller than for the hedonic double imputation index.

It is not surprising that the hedonic double imputation index performs poorly for the smallest amounts of data as a linear regression model demands a certain amount of data to produce a good and robust result. The period with the least amount of data for the 5% reduction only contains 18 data points and that gives a regression model with few degrees of freedom. One can clearly see that the time dummy method is more robust as the number of data points decrease. The reason for this is that the time dummy method still has a high number of degrees of freedom even for the smaller data sets which the double imputation method lacks.

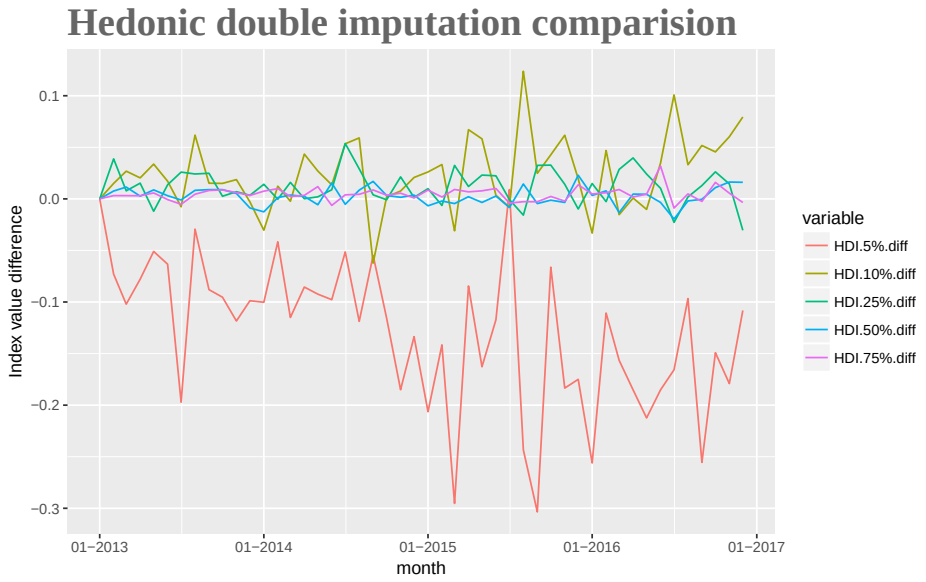


Figure 6.35: The difference in the Double Hedonic imputation index from model 2 with Huber regression methodology when 5% (HDI.5%.diff), 10% (HDI.10%.diff), 25% (HDI.25%.diff), 50% (HDI.50%.diff) and 75% (HDI.75%.diff) of the data from the Stockholm 2013-2016 data set is selected at random. The HDI.10%.diff index is modeled using OLS regression due to the confidence matrix lacking full rank in some time periods.

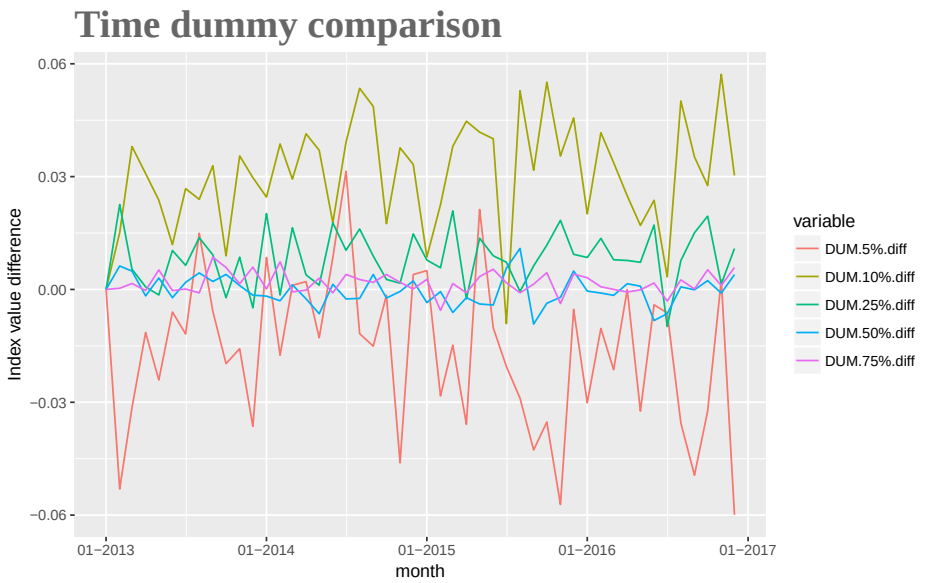


Figure 6.36: The difference in the hedonic time dummy index with Huber regression methodology when 5% (DUM.5%.diff), 10% (DUM.10%.diff), 25% (DUM.25%.diff), 50% (DUM.50%.diff) and 75% (DUM.75%.diff) of the data from the Stockholm 2013-2016 data set is selected at random.

7 Summary and Conclusions

In this section the results will be discussed and a recommendation of index modelling methodology for different settings will be provided.

7.1 Summary of results

The analysis and result section started by finding the best linear valuation model for the hedonic double imputation method. The analysis show that the apartment data has a tail heavy distribution and that a log-linear model provides better result than the regular linear model. The tail heaviness also suggests that a robust method for choosing the model should be used. Some analysis was done on the robust Huber regression model which provided similar results as the OLS model but with the benefit that the outliers had a smaller impact on the model. Analysing the residual vs fitted value plots one can see that the log-linear models provides good results in the middle region around the mean fitted value but that the residuals on the extreme values (both small and large) are not as good. This would have been a problem if one was interested in performing statistical tests on the coefficients in the model (as most test assumes that the errors follow a normal distribution), but we are only interested in the prediction capabilities of the linear model in this thesis. For modelling an index the mean characteristics of two periods is used and these characteristics will typically produce values which lie in the middle of the training range where the model performs very well, so the worse fit near the edges of the training interval is not a substantial problem.

All subset analysis in combination with cross validation of the best results was performed to find the best regression model for both the valuation model for the double imputation index and the time dummy index model. The methodology with all possible subset regressions finds the best models using all data and those models are then tested with cross validation to remove models that could suffer from an over fitting problem. One could perform cross validation on every possible model in the all subsets however this requires extensive computations and was therefore not performed. One could have chosen to perform a bootstrap method to validate the valuation models as well but with the rather large amount of data available a cross validation check is adequate.

The modeled indexes show that the method with dynamic choice of cluster location dummies produced the best results of the double imputation models. The reason for this is that more information could be included in the modeling of each point of the index. This raises the question if the valuation models used in the index should be even more dynamic and that the fitting parameters should be chosen in a more dynamic way in each period. That would have lead to a more complex index methodology which is not preferable. The more dynamic index methodology was not tested in this report due to the limitation of the scope. However this methodology will be mentioned in the suggestion for further studies section.

By plotting the average value of the underlying characteristics (independent variables in the valuation models, see figure (6.32) and (6.33)) one can see that the characteristics are relatively constant in the different time periods. This is not very surprising as a standard apartment does not change much during a time period of four years. This would indicate that the hedonic time dummy method and the hedonic double imputation method would produce similar results, which is also shown in figure (6.29). In the amount of data analysis we saw that the time dummy index performed better for the smaller dataset than the double imputation method as the difference between the index modeled using the full dataset and the stratified dataset was smaller for the time dummy index. This suggests that the time dummy methodology is more suitable for creation of a RPPI in a city where the transaction volume is not as big as in Stockholm. Another aspect that has not been mentioned earlier is the time period for which the index is modeled. If the time period is very long (many index values) the probability of a change in the underlying characteristics increases. A imputation approach would therefore be favourable as the index time period increases.

7.2 Suggestions for further studies

This thesis's main focus was to investigate the modelling of RPPIs using hedonic methods, mainly the hedonic double imputation index method and the hedonic time dummy index method. Different methods to model these indexes were investigated however due to the scope of this report these methods were not covered. We will therefore mention some interesting questions which were not analysed in this report but could be interesting areas for future investigations.

The geographical representation of the sold apartments in this thesis was handled using geographical dummy variables in a combination with continuous variables which indicated the distance to the closest water and ocean. One could think of different methods to include the geographical aspect of the valuation of an object, including creating an two-dimensional price function based on sold objects where each apartment is assigned a continuous value. Booli provided such a surface which was modeled using a kriging algorithm (read more about kriging in [16]). This methodology would then solve some of the problem with the loss of degrees of freedom as the inclusion of the geographical dummy variables included in the models used in this thesis.

The residential property price index literature mentions stratified hedonic indexes where the RPPIs are modeled for smaller regions than for an entire city, these RPPIs are then weighted together to create an index for the entire city. This would be an interesting methodology to test as our tested models have the assumption that the coefficients for the characteristics should be the same in different areas. The area difference is only handled by the location dummy variables and those are constant variables. The previous analysis done in this report would suggest that one could test stratified hedonic indexes using the time dummy model as the double imputation model would require too much data in each sub area.

The dataset used in this report did not include apartment IDs or any other way to determine with confidence if the same apartment were sold multiple times during the index modelling period. If one would have access to a dataset with those properties one could have performed a repeat sale validation of the modeled indexes by measuring which index method that best predicted the price change.

The best double imputation model was the model with dynamic updating of the number of location dummys which were included in each period (model 2). As mentioned in earlier sections it would be interesting to examine an double imputation method where more of the model is chosen dynamically in each period.

It would also be interesting to test non-linear valuation models in the double imputation index. One example would be to use an artificial neural network-model (ANN) for the valuation in each period. This model would be more complex but would have the possibility to produce better results. It would also be interesting to model a version of the time dummy model with help of an ANN, one could train the ANN with the time dummies as input parameters and then use input data from 2 periods and create an index in the same way as for an double imputation index.

7.3 Recommendations for index modelling

The hedonic double imputation approach is more favourable over the hedonic time dummy approach when there exists enough data to produce a good and robust valuation model for each time period. We would therefore recommend using the hedonic double imputation method for modelling of an RPPI for the large city regions in Sweden (Stockholm, Göteborg and Malmö) and using a hedonic time dummy methodology for the remaining regions where the transaction volume is smaller. However if one wishes to use the same methodology for all regions the hedonic time dummy methodology is preferred. We would also recommend to use a robust regression methodology when modelling the index as our analysis showed better results for robust methods.

8 References

[1] Berndt, E. R., and N. J. Rappaport. 2001. Price and quality of desktop and mobile personal computers: A quarter- century historical overview. American Economic Review 91 (2): 268– 73.

[2] Bloom mm, Sannolikhetsteori och statistikteori med tillämpningar, Fifth edition, Studentliteratur, 2005

[3] Bollen, Kenneth A.; Jackman, Robert W. (1990). Fox, John; Long, J. Scott, eds. Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. Modern Methods of Data Analysis. Newbury Park, CA: Sage. pp. 257–91.

[4] Chaitra H. Nagaraja, Lawrence D. Brown, Susan M. Wachter, House Price Index Methodology, U.S. Census Bureau, United States, 2010

[5] Croot, E 2010 "lecture notes for course 3225", Georgia Institute of Technology, School of Mathematics, http://people.math.gatech.edu/ ecroot/3225/maximum˙likelihood.pdf

[6] Dan Bradu and Yair Mundlak, Estimation in Lognormal Linear Models, Journal of the American Statistical Association, 1970

[7] David Birkes, Yadolah Dodge, Alternative Methods of Regression, First Edition, Wiley-Interscience Publication, 1993

[8] de Haan, J. (2010), "Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and Other Approaches", Journal of Economics and Statistics 230(6), 772-791.

[9] Diewert, W.E., S. Heravi and M. Silver (2009), "Hedonic Imputation Versus Time Dummy Hedonic Indexes", pp. 161-196 in Price Index Concepts and Measurement, W.E. Diewert, J. Greenlees and C. Hulten (eds.), NBER Studies in Income and Wealth, Chicago: University of Chicago Press.

[10] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, Introduction to Linear Regression Analysis, Fifth Edition, Wiley, 2012

[11] Eurostat, Handbook on Residential Property Prices Indices (RPPIs) 2013 edition, Publications Office of the European Union, Luxembourg, 2013

[12] Giovanni Seni, John F. Elder, Ensemble Methods in Data Mining: Improving AccuracyThrough Combining Predictions, Elder Research, 2010

[13] Hasintem ,T., and Qian, J., 2014, Glmnet Vignette, accessed 2017-04-13, https://web.stanford.edu/ hastie/glmnet/glmnet˙alpha.html

[14] Huber, P.J, Ronchetti, E.M, Robust Statistics, second edition, John Wiley & Sons, 2009

[15] Michy Alice,July 20, 2016,Performing Principal Components Regression (PCR) in R, accessed 29 Mars 2017, https://www.r-bloggers.com/performing-principal-components-regression-pcr-in-r/

[16] N.H. Bingham, John M. Fry, Regression - Linear Models in Statistics, First Edition, Springer, 2011

[17] Trevor Hastie, Robert Tibshirani and Jerome Friedman, The Elements of Statistical Learning, Second Edition, Springer, 2009

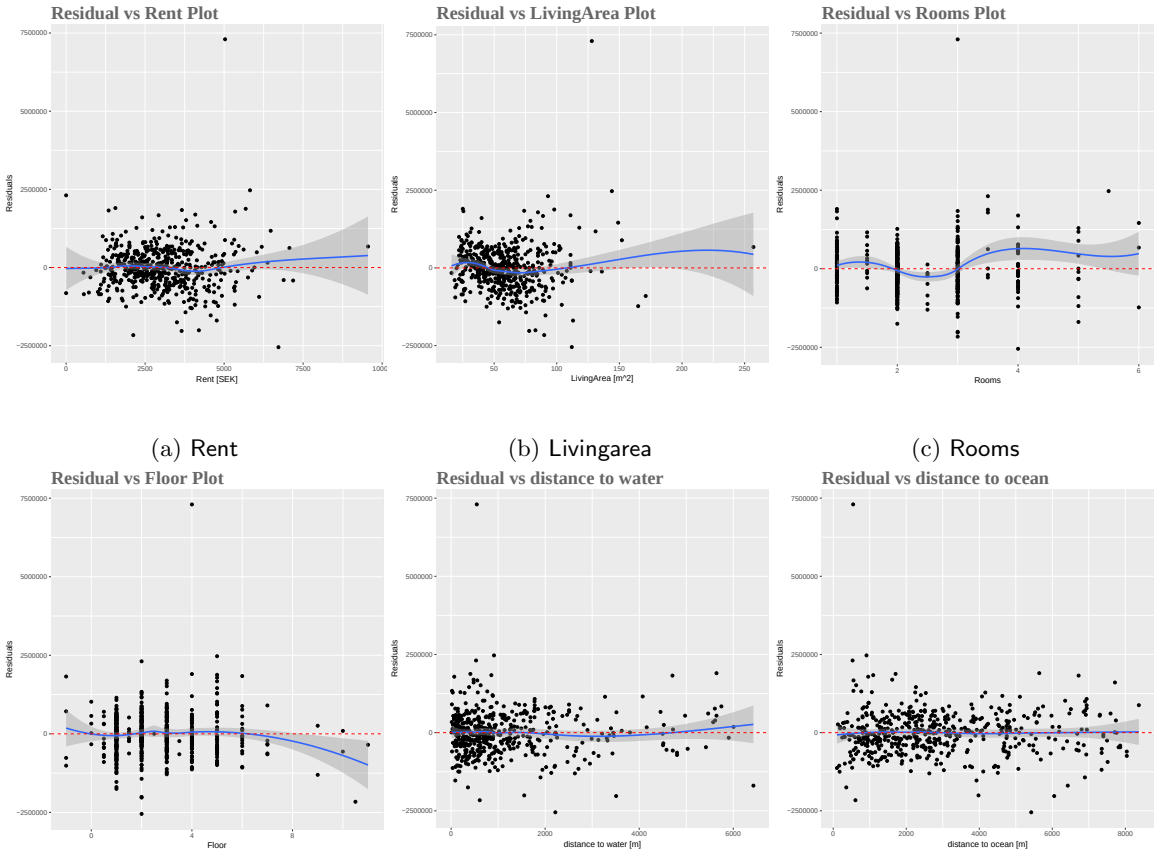
[18] Valueguard, Nasdaq OMX Valueguard-KTH Housing Index (HOX®) Methodology,Valuegurard Index Sweden AB, 2011, accessed 19-04-2017, http://www.valueguard.se/sites/default/files/HOX˙Methodology.pdf

[19] Valueguard, HOX Product Description,Valuegurard Index Sweden AB, 2011, accessed 19-04-2017, http://www.valueguard.se/sites/default/files/HOX˙Product˙Description.pdf

[20] W. Erwin Diewert, John S. Greenlees and Charles R. Hulten, Price Index Concepts and Measurement, University of Chicago Press, 2009, p 161 - 196

[21] Weisberg S, "handout robust regression", 2013, accessed 11-05-2017, University of Minnesota, "http://users.stat.umn.edu/ sandy/courses/8053/handouts/robust.pdf"

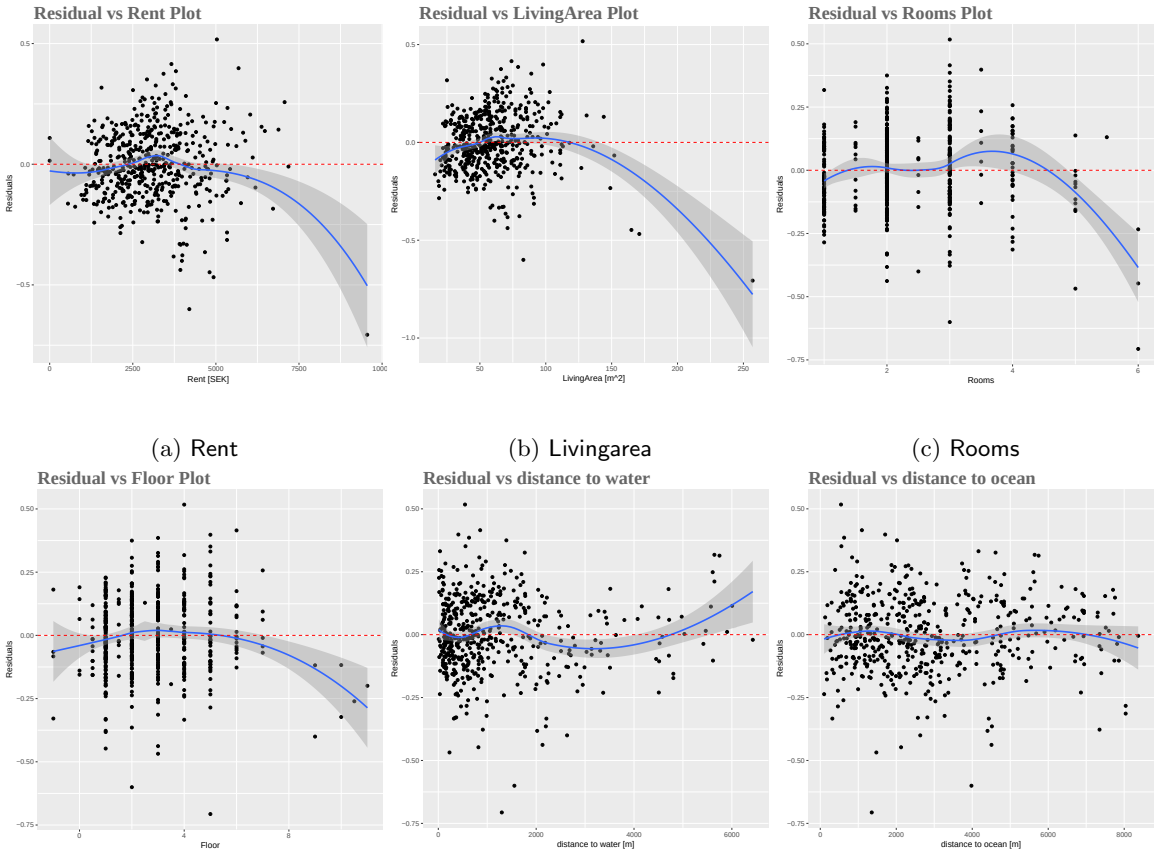
9 Appendix



(a) Rent (b) Livingarea (c) Rooms

(d) Floor (e) location.distance.water (f) location.distance.ocean

Figure 9.1: The residual plotted against the variables rent, livingArea, rooms, floor, location.distance.water and location.distance.ocean



(a) Rent (b) Livingarea (c) Rooms

(d) Floor (e) location.distance.water (f) location.distance.ocean

Figure 9.2: The residual plotted against the variables rent, livingArea, rooms, floor, location.distance.water and location.distance.ocean

Linear model for December 2016

Best models Adjusted R2

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.904	0.894	5.5E+11	15463.6	15680.1	1	1	1	1	0	0	1	1	1	1	1	1
0.904	0.893	5.5E+11	15465.4	15686.2	1	1	1	1	0	1	1	1	1	1	1	1
0.904	0.893	5.5E+11	15465.5	15686.3	1	1	1	1	1	0	1	1	1	1	1	1
0.903	0.893	5.5E+11	15463.9	15676.3	1	0	1	1	0	0	1	1	1	1	1	1
0.903	0.893	5.5E+11	15464.2	15676.5	1	1	1	1	0	0	1	1	0	1	1	1
0.904	0.893	5.5E+11	15467.4	15692.5	1	1	1	1	1	1	1	1	1	1	1	1
0.903	0.893	5.5E+11	15465.7	15682.3	1	0	1	1	0	1	1	1	1	1	1	1
0.903	0.893	5.5E+11	15465.8	15682.3	1	0	1	1	1	0	1	1	1	1	1	1
0.903	0.893	5.5E+11	15466.1	15682.7	1	1	1	1	0	1	1	1	0	1	1	1
0.903	0.893	5.5E+11	15466.1	15682.7	1	1	1	1	1	0	1	1	0	1	1	1

Best models Sum of squared residuals

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.904	0.894	5.5E+11	15463.6	15680.1	1	1	1	1	0	0	1	1	1	1	1	1
0.904	0.893	5.5E+11	15465.4	15686.2	1	1	1	1	0	1	1	1	1	1	1	1
0.904	0.893	5.5E+11	15465.5	15686.3	1	1	1	1	1	0	1	1	1	1	1	1
0.903	0.893	5.5E+11	15463.9	15676.3	1	0	1	1	0	0	1	1	1	1	1	1
0.903	0.893	5.5E+11	15464.2	15676.5	1	1	1	1	0	0	1	1	0	1	1	1
0.904	0.893	5.5E+11	15467.4	15692.5	1	1	1	1	1	1	1	1	1	1	1	1
0.903	0.893	5.5E+11	15465.7	15682.3	1	0	1	1	0	1	1	1	1	1	1	1
0.903	0.893	5.5E+11	15465.8	15682.3	1	0	1	1	1	0	1	1	1	1	1	1
0.903	0.893	5.5E+11	15466.1	15682.7	1	1	1	1	0	1	1	1	0	1	1	1
0.903	0.893	5.5E+11	15466.1	15682.7	1	1	1	1	1	0	1	1	0	1	1	1

Figure 9.3: The tables over all the best models due to the Adjusted R2 criterion and sum of squared residuals criterion for the linear model (part 1 of 2).

Linear model for December 2016

Best models AIC

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.904	0.894	5.5E+11	15463.6	15680.1	1	1	1	1	0	0	1	1	1	1	1	1
0.903	0.893	5.5E+11	15463.9	15676.3	1	0	1	1	0	0	1	1	1	1	1	1
0.903	0.893	5.5E+11	15464.2	15676.5	1	1	1	1	0	0	1	1	0	1	1	1
0.903	0.893	5.5E+11	15464.4	15672.4	1	0	1	1	0	0	1	1	0	1	1	1
0.903	0.893	5.5E+11	15465.4	15677.7	0	1	1	1	0	0	1	1	1	1	1	1
0.904	0.893	5.5E+11	15465.4	15686.2	1	1	1	1	0	1	1	1	1	1	1	1
0.904	0.893	5.5E+11	15465.5	15686.3	1	1	1	1	1	0	1	1	1	1	1	1
0.903	0.893	5.5E+11	15465.7	15682.3	1	0	1	1	0	1	1	1	1	1	1	1
0.903	0.893	5.5E+11	15465.8	15682.3	1	0	1	1	1	0	1	1	1	1	1	1
0.903	0.893	5.5E+11	15466.1	15682.7	1	1	1	1	0	1	1	1	0	1	1	1

Best models BIC

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.872	0.868	6.8E+11	15542.0	15610.0	1	0	1	1	0	0	1	1	1	1	1	0
0.869	0.866	6.9E+11	15550.0	15613.7	1	0	1	1	0	0	1	0	1	1	1	0
0.872	0.868	6.8E+11	15543.5	15615.7	1	0	1	1	1	0	1	1	1	1	1	0
0.872	0.868	6.8E+11	15543.7	15615.9	1	1	1	1	0	0	1	1	1	1	1	0
0.872	0.868	6.8E+11	15543.8	15615.9	1	0	1	1	0	1	1	1	1	1	1	0
0.869	0.866	7E+11	15551.7	15619.7	1	0	1	1	1	0	1	0	1	1	1	0
0.869	0.866	7E+11	15551.8	15619.7	1	1	1	1	0	0	1	0	1	1	1	0
0.869	0.866	7E+11	15551.9	15619.8	1	0	1	1	0	1	1	0	1	1	1	0
0.872	0.868	6.8E+11	15545.1	15621.6	1	1	1	1	1	0	1	1	1	1	1	0
0.872	0.868	6.8E+11	15545.3	15621.8	1	0	1	1	1	1	1	1	1	1	1	0

Figure 9.4: The tables over all the best models due to the AIC criterion and BIC criterion for the linear model (part 2 of 2).

	VIF		VIF
rent	27.49	cluster.30dummy	2.40
rent.2	27.13	cluster.31dummy	3.49
livingArea	33.00	cluster.33dummy	3.28
livingArea.2	23.11	cluster.34dummy	3.13
rooms	36.36	cluster.35dummy	4.93
rooms.2	30.26	cluster.36dummy	5.38
floor	7.88	cluster.39dummy	2.78
floor.2	7.65	cluster.44dummy	2.61
location.distance.ocean	17.07	cluster.45dummy	6.71
funkis.CT.dummy	2.40	cluster.47dummy	2.63
folkhem.CT.dummy	1.43	cluster.56dummy	3.77
miljonprogram.CT.dummy	1.25	cluster.67dummy	3.13
osubventionerat.CT.dummy	1.53	cluster.68dummy	1.79
mordern.CT.dummy	1.51	cluster.71dummy	2.61
nyproduktion.CT.dummy	1.86	cluster.72dummy	1.93
missing.CT.dummy	1.60	cluster.77dummy	2.97
cluster.1dummy	3.42	cluster.81dummy	2.94
cluster.2dummy	4.27	cluster.86dummy	2.72
cluster.10dummy	4.54	cluster.89dummy	2.25
cluster.11dummy	3.69	cluster.96dummy	2.79
cluster.13dummy	3.33	cluster.97dummy	5.64
cluster.15dummy	1.66	cluster.98dummy	3.84
cluster.23dummy	3.80	cluster.101dummy	2.24
cluster.24dummy	5.61	cluster.102dummy	8.21
cluster.25dummy	5.16	cluster.104dummy	1.93

Table 9.1: The VIF values for equation (6.3). VIF values larger than 10 is marked with red

Log-linear model for December 2016

Best models Adjusted R2

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.929	0.921	0.0135	-706.9	-486.1	1	0	1	1	1	1	1	1	1	1	1	1
0.929	0.921	0.0135	-705.9	-480.9	1	1	1	1	1	1	1	1	1	1	1	1
0.929	0.921	0.0135	-706.6	-485.8	1	1	1	1	1	1	1	1	0	1	1	1
0.929	0.921	0.01351	-707.4	-490.9	1	0	1	1	1	1	1	1	0	1	1	1
0.928	0.920	0.01361	-703.5	-486.9	1	0	1	1	1	0	1	1	1	1	1	1
0.928	0.920	0.01361	-702.4	-481.6	1	1	1	1	1	0	1	1	1	1	1	1
0.928	0.920	0.01363	-703.7	-491.4	1	0	1	1	1	0	1	1	0	1	1	1
0.928	0.920	0.01363	-702.8	-486.2	1	1	1	1	1	0	1	1	0	1	1	1
0.927	0.920	0.01376	-697.1	-476.3	1	1	1	1	0	1	1	1	1	1	1	1
0.927	0.920	0.01376	-698.0	-481.4	1	0	1	1	0	1	1	1	1	1	1	1

Best models Sum of squared residuals

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.929	0.921	0.0135	-706.9	-486.1	1	0	1	1	1	1	1	1	1	1	1	1
0.929	0.921	0.0135	-705.9	-480.9	1	1	1	1	1	1	1	1	1	1	1	1
0.929	0.921	0.0135	-706.6	-485.8	1	1	1	1	1	1	1	1	0	1	1	1
0.929	0.921	0.01351	-707.4	-490.9	1	0	1	1	1	1	1	1	0	1	1	1
0.928	0.920	0.01361	-703.5	-486.9	1	0	1	1	1	0	1	1	1	1	1	1
0.928	0.920	0.01361	-702.4	-481.6	1	1	1	1	1	0	1	1	1	1	1	1
0.928	0.920	0.01363	-703.7	-491.4	1	0	1	1	1	0	1	1	0	1	1	1
0.928	0.920	0.01363	-702.8	-486.2	1	1	1	1	1	0	1	1	0	1	1	1
0.927	0.920	0.01376	-697.1	-476.3	1	1	1	1	0	1	1	1	1	1	1	1
0.927	0.920	0.01376	-698.0	-481.4	1	0	1	1	0	1	1	1	1	1	1	1

Figure 9.5: The tables over all the best models due to the Adjusted R2 criterion and sum of squared residuals criterion for the loglinear model (part 1 of 2).

Log-linear model for December 2016

Best models AIC

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.929	0.921	0.01351	-707.4	-490.9	1	0	1	1	1	1	1	1	0	1	1	1
0.929	0.921	0.0135	-706.9	-486.1	1	0	1	1	1	1	1	1	1	1	1	1
0.929	0.921	0.0135	-706.6	-485.8	1	1	1	1	1	1	1	1	0	1	1	1
0.929	0.921	0.0135	-705.9	-480.9	1	1	1	1	1	1	1	1	1	1	1	1
0.928	0.920	0.01363	-703.7	-491.4	1	0	1	1	1	0	1	1	0	1	1	1
0.928	0.920	0.01361	-703.5	-486.9	1	0	1	1	1	0	1	1	1	1	1	1
0.928	0.920	0.01363	-702.8	-486.2	1	1	1	1	1	0	1	1	0	1	1	1
0.928	0.920	0.01361	-702.4	-481.6	1	1	1	1	1	0	1	1	1	1	1	1
0.927	0.920	0.01376	-698.0	-481.4	1	0	1	1	0	1	1	1	1	1	1	1
0.927	0.919	0.01378	-698.0	-485.7	1	0	1	1	0	1	1	1	0	1	1	1

Best models BIC

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.928	0.920	0.01363	-703.7	-491.4	1	0	1	1	1	0	1	1	0	1	1	1
0.929	0.921	0.01351	-707.4	-490.9	1	0	1	1	1	1	1	1	0	1	1	1
0.927	0.919	0.01383	-697.1	-489.0	1	0	1	1	0	0	1	1	0	1	1	1
0.928	0.920	0.01361	-703.5	-486.9	1	0	1	1	1	0	1	1	1	1	1	1
0.928	0.920	0.01363	-702.8	-486.2	1	1	1	1	1	0	1	1	0	1	1	1
0.929	0.921	0.0135	-706.9	-486.1	1	0	1	1	1	1	1	1	1	1	1	1
0.929	0.921	0.0135	-706.6	-485.8	1	1	1	1	1	1	1	1	0	1	1	1
0.927	0.919	0.01378	-698.0	-485.7	1	0	1	1	0	1	1	1	0	1	1	1
0.927	0.919	0.0138	-697.2	-484.9	1	0	1	1	0	0	1	1	1	1	1	1
0.927	0.919	0.01382	-696.6	-484.3	1	1	1	1	0	0	1	1	0	1	1	1

Figure 9.6: The tables over all the best models due to the AIC criterion and BIC criterion for the log-linear model (part 2 of 2).

Time dummy model

Best models Adjusted R2

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.914	0.913	0.01762	-67451	-66406	1	1	1	1	1	1	1	1	1	1	1	1
0.914	0.913	0.01764	-67402	-66365	1	1	1	1	1	1	1	1	0	1	1	1
0.913	0.913	0.01764	-67392	-66355	1	0	1	1	1	1	1	1	1	1	1	1
0.913	0.913	0.01766	-67341	-66314	1	0	1	1	1	1	1	1	0	1	1	1
0.913	0.913	0.0177	-67200	-66163	1	1	1	1	1	0	1	1	1	1	1	1
0.913	0.913	0.01772	-67153	-66125	1	0	1	1	1	0	1	1	1	1	1	1
0.913	0.913	0.01772	-67152	-66124	1	1	1	1	1	0	1	1	0	1	1	1
0.913	0.913	0.01773	-67122	-66085	1	1	1	1	1	1	1	0	1	1	1	1
0.913	0.913	0.01773	-67104	-66085	1	0	1	1	1	0	1	1	0	1	1	1
0.913	0.913	0.01774	-67073	-66045	1	1	1	1	1	1	1	0	0	1	1	1

Best models Sum of squared residuals

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.914	0.913	0.01762	-67451	-66406	1	1	1	1	1	1	1	1	1	1	1	1
0.914	0.913	0.01764	-67402	-66365	1	1	1	1	1	1	1	1	0	1	1	1
0.913	0.913	0.01764	-67392	-66355	1	0	1	1	1	1	1	1	1	1	1	1
0.913	0.913	0.01766	-67341	-66314	1	0	1	1	1	1	1	1	0	1	1	1
0.913	0.913	0.0177	-67200	-66163	1	1	1	1	1	0	1	1	1	1	1	1
0.913	0.913	0.01772	-67153	-66125	1	0	1	1	1	0	1	1	1	1	1	1
0.913	0.913	0.01772	-67152	-66124	1	1	1	1	1	0	1	1	0	1	1	1
0.913	0.913	0.01773	-67122	-66085	1	1	1	1	1	1	1	0	1	1	1	1
0.913	0.913	0.01773	-67104	-66085	1	0	1	1	1	0	1	1	0	1	1	1
0.913	0.913	0.01774	-67073	-66045	1	1	1	1	1	1	1	0	0	1	1	1

Figure 9.7: The tables over all the best models due to the Adjusted R2 criterion and sum of squared residuals criterion for the time dummy model (part 1 of 2).

Time dummy model

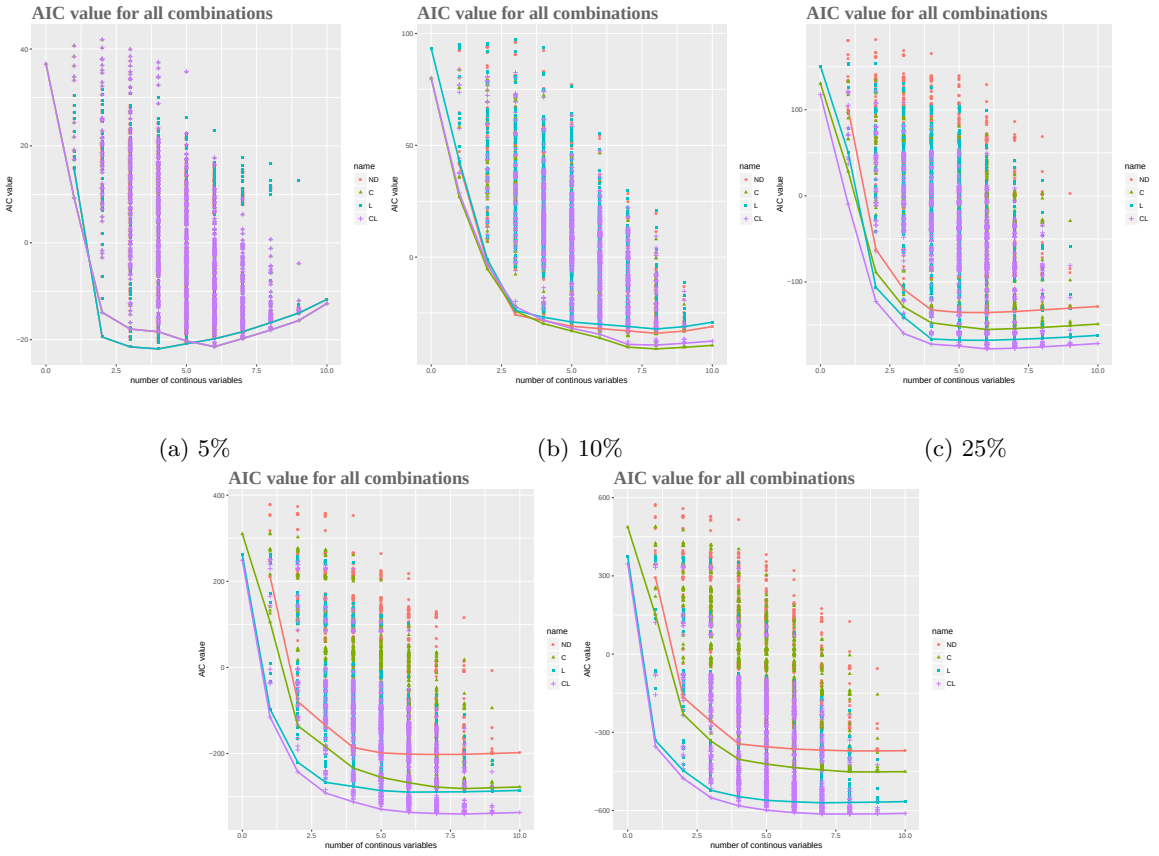
Best models AIC

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.914	0.913	0.01762	-67451	-66406	1	1	1	1	1	1	1	1	1	1	1	1
0.914	0.913	0.01764	-67402	-66365	1	1	1	1	1	1	1	1	0	1	1	1
0.913	0.913	0.01764	-67392	-66355	1	0	1	1	1	1	1	1	1	1	1	1
0.913	0.913	0.01766	-67341	-66314	1	0	1	1	1	1	1	1	0	1	1	1
0.913	0.913	0.0177	-67200	-66163	1	1	1	1	1	0	1	1	1	1	1	1
0.913	0.913	0.01772	-67153	-66125	1	0	1	1	1	0	1	1	1	1	1	1
0.913	0.913	0.01772	-67152	-66124	1	1	1	1	1	0	1	1	0	1	1	1
0.913	0.913	0.01773	-67122	-66085	1	1	1	1	1	1	1	0	1	1	1	1
0.913	0.913	0.01773	-67104	-66085	1	0	1	1	1	0	1	1	0	1	1	1
0.913	0.913	0.01774	-67073	-66045	1	1	1	1	1	1	1	0	0	1	1	1

Best models BIC

R2	adj_R2	SSres	AIC	BIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
0.914	0.913	0.01762	-67451	-66406	1	1	1	1	1	1	1	1	1	1	1	1
0.914	0.913	0.01764	-67402	-66365	1	1	1	1	1	1	1	1	0	1	1	1
0.913	0.913	0.01764	-67392	-66355	1	0	1	1	1	1	1	1	1	1	1	1
0.913	0.913	0.01766	-67341	-66314	1	0	1	1	1	1	1	1	0	1	1	1
0.913	0.913	0.0177	-67200	-66163	1	1	1	1	1	0	1	1	1	1	1	1
0.913	0.913	0.01772	-67153	-66125	1	0	1	1	1	0	1	1	1	1	1	1
0.913	0.913	0.01772	-67152	-66124	1	1	1	1	1	0	1	1	0	1	1	1
0.913	0.913	0.01773	-67122	-66085	1	1	1	1	1	1	1	0	1	1	1	1
0.913	0.913	0.01773	-67104	-66085	1	0	1	1	1	0	1	1	0	1	1	1
0.913	0.913	0.01774	-67073	-66045	1	1	1	1	1	1	1	0	0	1	1	1

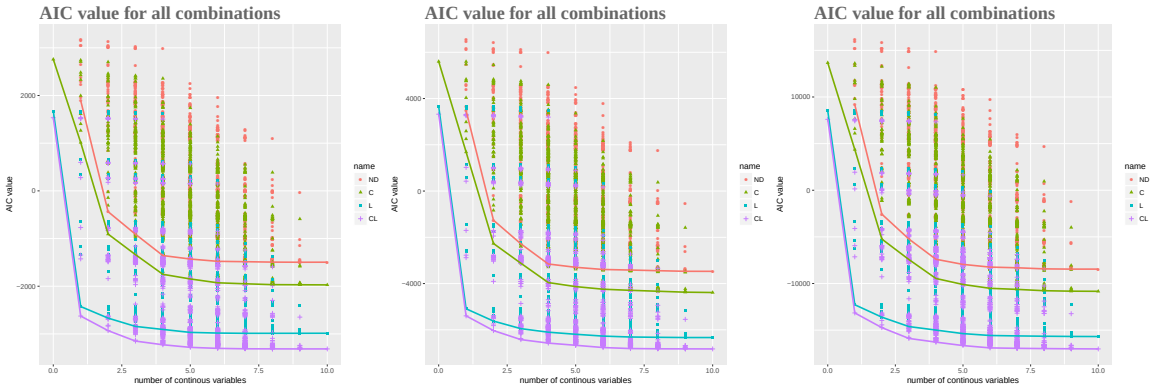
Figure 9.8: The tables over all the best models due to the AIC criterion and BIC criterion for the time dummy model (part 1 of 2).



(a) 5% (b) 10% (c) 25%

(d) 50% (e) 75%

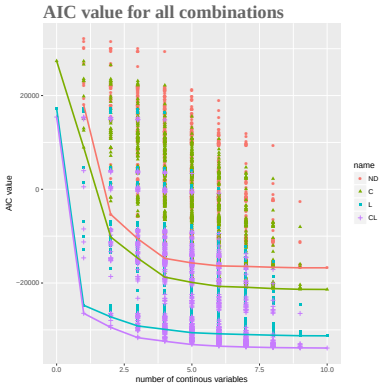
Figure 9.9: The AIC plots for the all subset analysis for the hedonic double imputation model used to compare different amount of data. The best model for each dataset can be seen in figure (9.11).



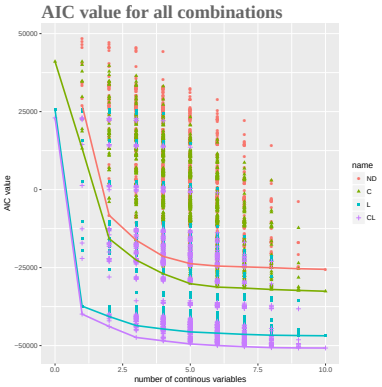
(a) 5%

(b) 10%

(c) 25%



(d) 50%



(e) 75%

Figure 9.10: The AIC plots for the all subset analysis for the time dummy model used to compare different amount of data. The best model for each dataset can be seen in figure (9.11).

	AIC	rent	rent.2	livingArea	livingArea.2	rooms	rooms.2	floor	floor.2	location.distance.water	location.distance.ocean	Construction.fam	Location.fam
data5HDI	-21.9047	0	1	1	0	0	0	0	1	0	1	0	0
data10HDI	-41.0449	1	0	1	1	0	1	1	1	1	1	1	0
data25HDI	-177.794	0	1	1	1	0	0	1	0	1	1	1	1
data50HDI	-340.405	1	0	1	1	0	1	1	1	1	1	1	1
data75HDI	-613.316	0	1	1	1	1	0	1	1	0	1	1	1
data5DUM	-3315	1	1	1	1	1	1	1	0	1	1	1	1
data10DUM	-6826.38	1	1	1	1	1	1	1	1	1	1	1	1
data25DUM	-17015.2	1	1	1	1	1	1	1	1	1	1	1	1
data50DUM	-33876.3	1	1	1	1	1	1	1	1	1	1	1	1
data75DUM	-50779.8	1	1	1	1	1	1	1	1	1	1	1	1

Figure 9.11: The best models from the data amount analysis.