Stockholms universitet

Construction of Price Indices for Stockholm Condominiums

Simon Melamed

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2020:14
Matematisk statistik
Juni 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2020:14**
http://www.math.su.se

Construction of Price Indices for Stockholm Condominiums

Simon Melamed*

June 2020

Abstract

The main aim of this thesis is to explore the world of house price indices and the statistical mechanics behind them. I introduce two commonly used models for price indices around the world; the hedonic time dummy model and the repeat sales model. In Section 3, six different price indices based on transactional data on Stockholm's condominiums are created. These indices spanned over the period from January 2014 to December 2018. For the construction of these indices, statistical regression methods were required to estimate the coefficients of the price models. The regression methods used in this thesis are the Ordinary Least Squares, Weighted Least Squares, and Robust regression. Lastly, I used different validation metrics to revise the indices validity and to compare them with each other. All indices constructed in this thesis illustrated similar price movement patterns, an upward trend from January 2014 towards the end of 2017 where the peak turned and the price indices dropped by approximately 15%.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: simon@melamed.se. Supervisor: Kristoffer Lindensjö.

Acknowledgements

The following is a Bachelor's thesis (of 15 ECTS) in Mathematics and Statistics at Stockholm University. I want to start by thanking my supervisor, Kristoffer Lindensjö, for his support during the writing process. Additionally, I want to thank Hans Flink at Mäklarstatistik AB for providing the data which made this thesis feasible. Lastly, I want to thank my dear family and friends for their support during this entire time.

Contents

1 Introduction

In Stockholm, Sweden, the housing market oftentimes seems to be at the center of attention for many people. Should I invest in a new home now? Is there a market bubble that will burst any minute and cause the prices of condominiums to plummet? I believe it is fair to say that deciding to invest in property is a big personal risk for many people. Luckily, there are residential property price indices which can be useful tools to help navigate this life-changing transformation.

To use price indices for personal decision making is merely one of its many important uses. For example it is also used as a macro-economic indicator. It has been shown that there often is a high correlation between the state of the housing market and the economy (Goodhart and Hofmann, 2006, as cited by Eurostat et al, 2013, p. 17).

In this thesis we will discuss two common types of price indices, the *hedonic time dummy index* and the *repeat sales index*. The aim is to learn more about price index methodology and to produce indices for the condominium market in Stockholm.

2 Theory

2.1 Hedonic Regression

Economist Sherwin Rosen first introduced the theoretical framework for the hedonic price model for housing in 1974. According to his definition of hedonic prices, the hedonic price model assumes that goods can be priced depending on their characteristics. Therefore, prices can be looked upon as a bundle of characteristics. Each one of the inherent characteristics has its contributory price to the full price (Rosen, 1917, p. 35). For example, the price of my apartment can be viewed as the package deal of its characteristics such as location, living area, closeness to amenities and structure. Since these characteristics are not purchased separately but rather in a bundle, the hedonic price model can be used as a method to compute estimates of the contributory price of each one of these characteristics (Eurostat et al, 2013, p. 13). With these estimates, the willingness to pay for different characteristics of properties can be evaluated. In this thesis however, we will restrict ourselves to use the hedonic price model to construct hedonic price indices for condominiums.

2.1.1 Derivation of the Hedonic Model

Let us consider a hedonic price model for condominiums in Stockholm. We start by introducing the hedonic model with the assumption that the price

of the condominium, p_n^t is a function of some number, K, characteristics with quantities z_{nk}^t,

$$p_n^t = f(z_{n1}^t, ..., z_{nK}^t, \epsilon_n^t) \tag{1}$$

for $n \in \{1, 2, ..., N\}$ and $t \in \{0, 1, .., T\}$. Where n denotes the identification number for the condominium, ϵ_n^t is the random error term, and t denotes the period in which the condominium is sold. The two most well known hedonic specifications are the full linear model and the logarithmic linear model (Eurostat et al, 2013, p. 50).

The linear model is a standard multilinear regression model with the characteristics as independent covariates and the price as the dependent variable,

$$p_n^t = \alpha^t + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \epsilon_n^t \tag{2}$$

for $n \in \{1, 2, ..., N\}$ and $t \in \{0, 1, .., T\}$. In contrast, the logarithmic linear model has the natural logarithm of the price as the dependent variable,

$$\ln p_n^t = \alpha^t + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \epsilon_n^t. \tag{3}$$

In these hedonic specifications, (2) and (3), α^t is the intercept and β_k^t are the coefficients for the characteristics. Examples of characteristics are the number of rooms, living area, and monthly rent of the condominiums. These three characteristics will have different quantities, z_{nk}^t, for different condominiums in different periods. The parameters, α^t and β_k^t, may vary over time and they are estimated for each period $t \in \{0, 1, ..., T\}$ using some regression method. This is an appropriate model for the reality since the housing market conditions determine the marginal contributions of each covariate (Pakes, 2003, p. 6). For example, if the demand in the housing market is low, the coefficients β_k^t will tend to be lower. Because people are less willing to pay for a condominium and implicitly less willing to pay for the inherent characteristics of the condominium.

From these models, (2) and (3), the intercept terms α^t are used to construct a price index for each period t. In practice, using these models can be quite tedious since the regression has to be run for each period. Consequently, each regression is run on a smaller sample, potentially leading to inaccurate estimations. To escape this practical issue, one can make the assumption that β_k^t remain remotely constant when considering a short time interval. In this case we have the constrained version of (3) where $\beta_k^t = \beta_k$ which yield,

$$\ln p_n^t = \alpha^t + \sum_{k=1}^{K} \beta_k z_{nk}^t + \epsilon_n^t. \tag{4}$$

When using the constrained model, so-called *time dummy variables* are included to indicate which period each observation is from,

$$\ln p_n^t = \alpha^t + \sum_{k=1}^{K} \beta_k z_{nk}^t + \sum_{\tau=1}^{T} \gamma^\tau D_n^\tau + \epsilon_n^t \tag{5}$$

for $n \in \{1, ..., N\}$ and $t \in \{1, .., T\}$. For this specification, D_n^τ denotes a categorical variable that take the values,

$$D_n^\tau = \begin{cases} 1, & \text{if } \tau \text{ is the period of the sale,} \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

Furthermore, γ^τ is the coefficient for each time dummy variable. Note that τ ranges from 1 to T since 0 is put to base period to prevent perfect collinearity in the model. Perfect collinearity is when an explanatory variable is an exact linear function of the other explanatory variables and is therefore redundant to the model (Dormann et al, 2012, p. 28).

All of the steps to retrieve the hedonic time dummy model can be found in the Handbook on Residential Property Prices Indices (Eurostat et al, 2013, p. 50).

2.1.2 Construction of Hedonic Price index

From model (5) we will construct a simple price index using the estimated time dummy coefficients $\hat{\gamma}^\tau$ from the chosen regression. These estimates can be interpreted as a measure of how the period affects the logarithm of prices of condominiums. Therefore it is quite intuitive to use these estimates to construct the index. We will also have an estimate for the intercept, $\hat{\alpha}_0$, and estimates of the coefficients of the characteristics, $\hat{\beta}_k$. However, for our purposes, these estimates are merely used to factor out biases from our index and to guarantee a quality-adjusted index. Meaning, when we exponentiate the coefficients $\hat{\gamma}^\tau$ for the time dummy variables and multiply them with a 100, we have controlled for the inherent characteristics of the condominiums and receive a quality-adjusted price index with base period 0 and period τ,

$$P_{TD}^{0,\tau} = e^{\hat{\gamma}^\tau} \cdot 100.$$

Since we have put the base period to 0, the index will start at 100 and then go to $P_{TD}^{0,1} = e^{\hat{\gamma}^1} \cdot 100$ in period 1 and so forth up until period T which is latest period.

It should be emphasized that this is merely one version of hedonic regression. Eurostat et al (2013, p. 39) list the three most frequently used hedonic

regressions to be characteristics, imputation, and time dummy approaches. In this thesis we have only covered the time dummy hedonic specification.

2.1.3 Drawbacks of the Hedonic Time Dummy method

To construct price indices using the hedonic time dummy method is widely practiced across the world due to its simplicity and straightforwardness. For example, Valueguard Index Sweden AB releases monthly price indices on housing in Sweden as a whole and different large cities such as Stockholm and Gothenburg, using the hedonic time dummy price model (HOX Sverige, 2020). However, there are some drawbacks to this indexing methodology.

Firstly, there is a problem of omitted variable bias. There might be some underlying factor that affects the price of the house that is not being accounted for in the model specification. Perhaps the condominium is near to some disturbance, such as roads or trash disposals, which in turn reflects on the price but is not accounted for in the hedonic model. The causing effect of the omitted variable bias is that the error term and standard deviations of the estimates become greater and the accuracy of the index smaller (Eurostat et al, 2013, p. 51).

Secondly, there is a problem of having the assumption that the coefficients of the characteristics remain constant over time, since this is not true in reality (Pakes, 2003, p. 6). Since the coefficients β_k will tend to be lower when the demand in the market is lower and higher when the opposite is true, as mentioned previously.

With the assumption of constant coefficients comes an additional problem of revision. As time passes and more data is collected, the time dummy method requires the previous index points to be revised and recalculated. Since we run a new regression containing new data and estimate new coefficients. The solution to this issue is normally to run regression on adjacent periods, $t-1$ and t, and then multiply them together (Eurostat, 2013, p. 52). In result, we have a time series that is not affected by the revision instability issue.

Lastly, when it comes to index construction there is always the risk of sample selection bias (Eurostat et al, 2013, p. 67). The transactions at a specific period may not necessarily reflect the housing market in its entirety.

2.2 Repeat Sales Method

In this section we will discuss another kind of approach, called the *repeat sales method*, which was introduced in 1963 by Bailey, Muth, and Nourse (Bailey et al. 1963). For this method less information of the transactions is required in comparison to the hedonic approach. The only components needed is contract price, sale date, and property identification. This is because we restrict ourselves to consider condominiums that have been sold

multiple times during the time interval of interest. In result, we have data on multiple transactions for the same property and hence a direct measure of how prices for condominiums change over time.

Indices constructed using this method is also commonly used all over the world. In the US, the S&P/Case-Shiller Housing index tracks changes in home prices nationwide and is constructed using the repeat sales method (S&P Dow Jones Indices LLC, 2020).

We derive the model for the repeat sales method from the starting point of the constrained log-linear hedonic model, see equation (4). However, since we exclusively consider condominiums that are sold multiple times during the time interval, we assume that the amounts of the characteristics are constant during the time interval. For example, we assume that a condominium that has 3 number of rooms, 50 square meters of living area, and a monthly fee of 1500 Swedish crowns at the period of the first sale, has the same amounts of these characteristics at the period of the second sale. Generally, this implies that $z_{nk}^t = z_{nk}$ for all condominium characteristics. Which yields, continuing from equation (4),

$$\ln\left(p_n^t\right) = \alpha^t + \sum_{k=1}^{K} \beta_k z_{nk} + \epsilon_n^t. \tag{7}$$

A model of the difference of the logarithm of the price of condominium n, with first sale at period s and resale at period t, such that $0 \leq s < t \leq T$, is the given by,

$$ln(p_n^t) - ln(p_n^s) = \alpha^t + \sum_{k=1}^{K} \beta_k z_{nk} + \epsilon_n^t - (\alpha^s + \sum_{k=1}^{K} \beta_k z_{nk} + \epsilon_n^s),$$

note that some of the terms cancel out and we retrieve,

$$ln(p_n^t) - ln(p_n^s) = \alpha^t - \alpha^s + \epsilon_n^t - \epsilon_n^s. \tag{8}$$

Note that the assumption of this model is that the change of the condominium prices is the practically the same for all condominiums n, except for an error term, $\epsilon_n^t - \epsilon_n^s$. Since, $\alpha^t - \alpha^s$ does not depend on n. From (8), we arrive at the repeat sales equation,

$$ln(p_n^t) - ln(p_n^s) = \sum_{t=0}^{T} \delta^t D_n^t + \mu_n^t, \text{ for } n \in \{1, 2, ..., N\}. \tag{9}$$

Where D_n^t is a dummy variable taking the values,

$$D_n^t = \begin{cases} -1, & \text{if } t \text{ is the period of the first sale} \\ 1, & \text{if } t \text{ is the period of the resale} \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

for each condominium n in the repeat sales data set. Similarily as the hedonic case, μ_n^t denotes a random error term. The coefficients for the dummy variables in this model, δ^t will then be estimated and used to create the repeat sales indices.

Let us consider an example, suppose that the number of the dwelling is $n = 10$, and is first sold in period $t = 3$ and sold again in period $t = 30$. The repeat sales equation will then be,

$$ln(p_{10}^{30}) - ln(p_{10}^{3}) = \sum_{t=0}^{T} \delta^t D_{10}^t + \mu_{10}^{30} = \delta^3 \cdot -1 + \delta^{30} \cdot 1 + \mu_{10}^{30}. \tag{11}$$

On the left-hand side of (11) we have the difference of the logarithm of the prices of dwelling $n = 10$ in the two periods, which is known from the data. On the very right-hand side of (11) we have the time dummy coefficient for period $t = 30$ minus the time dummy coefficient for period $t = 3$ plus a random error term. When we use a regression to estimate the parameters for δ^{30} and δ^3, we choose the estimates that best fit our repeat sales data set. Hopefully, this example clarifies how the repeat sales equation works.

When there are more than two transactions of the same property in the repeat sales data set, all of the possible transaction pairs are utilized in the regression. For example, suppose that the property in the example is sold at three periods $t = \{3, 30, 42\}$, then all transaction pairs $3-30$, $3-42$, and $30-42$ are used in the model to estimate the coefficients.

2.2.1 Construction of Repeat Sales Index

To create an index using the repeat sales method, one has to run a regression, exactly as in the hedonic time dummy case, on the pooled repeat sales data. The data being pooled simply implies that we consider the data set containing observations from all the periods rather than regressing on samples from each period (Eurostat et al, 2013, p. 26). The repeat sales index is then calculated using the estimates of the coefficients received from the regression,

$$P_{RS}^{0,t} = e^{\hat{\delta}^t} \cdot 100. \tag{12}$$

Again, the index at the base period, $t = 0$, is equal to 100 since $e^{\hat{\delta}^0} = 1$. The estimates of the time dummy coeffiecients, $\hat{\delta}^t$, can be interpreted as a measure of condominium prices at period t (Bourassa et al, 2004, p. 5). For example, let us suppose that the prices of condominiums is higher in period $t = 10$ compared to another period $t = 20$. Consequently, the estimate $\hat{\delta}^{10}$ will be greater than $\hat{\delta}^{20}$. In result, the repeat sales index will reflect this difference and the index for period 10, $P_{RS}^{0,10} = e^{\hat{\delta}^{10}} \cdot 100$, will be greater

than the index for period 20, $P_{RS}^{0,20} = e^{\hat{\delta}^{20}} \cdot 100$. Therefore, the repeat sales index can be used to analyze the strength of the housing market in different periods of time.

The fact that the repeat sales index merely requires information on the contract price, transaction date, and property id, explains its popularity in housing literature and real estate agencies. Additionally, the repeat sales model is fully quality-adjusted, since we compare condominiums with themselves at two points in time (Bourassa et al, 2004, p. 3).

2.2.2 Drawbacks of the Repeat Sales Index

Expectedly, the repeat sales model has its drawbacks as well. One problem with this model is that it assumes that the characteristics, z_{nk}, are held constant over time, which is not necessarily true in all cases. The model does not take into account that there is a possibility that the condominium has been renovated between the time of the sales, which potentially would cause an increase in contract price. The model then assumes that the increase is a reflection of the housing stock and is wrongly reflected in the index. On the other hand, the repeat sales model does not account for the potential depreciation of condominiums either. However, the issue of appreciation has been counteracted by setting the minimum time interval between two sales to one year (Clapp and Giacotto, 1999 as cited in Clapham et al, 2004, p. 18). Thus, the risk of having home flips is minimized; buying and reselling property quickly does not normally obey the assumption of the constant characteristics due to renovations.

Another issue with this model is that we consider smaller data samples since we restrain ourselves to consider resales. This can lead to smaller data sets and fewer degrees of freedom for the regression, along with sample selection bias (Eurostat, 2013, p. 68). In result, this could imply unstable estimates of the coefficients and in turn an inaccurate index.

Similarily as the hedonic time dummy method, this method is also subject to the revision issue, since the regression is run on the pooled repeat sales data.

2.3 Regression Methods

In this section we will discuss the three regression methods; Ordinary Least Squares, Weighted Least Squares, and Robust regression. Later in the thesis, each method will be used and yield different indices. Theory in this section is relying heavily on the literature, Notes in Econometrics (2019).

2.3.1 Ordinary Least Squares

We will start with the most common regression method, Ordinary Least Squares. First, let us introduce some notations, the response variable vector

contains the data which we want to model,

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}.$$

In the hedonic time dummy model, the response variable vector contains the logarithmic prices of condominiums in different periods, $ln(p_n^t)$. Whereas, in the repeat sales approach, the response variable vector contains the differences of logarithmic prices of condominiums at two time points, $ln(p_n^t) - ln(p_n^s)$, where $0 \leq s < t \leq T$.

Then we have the parameter vector,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Which contains the coeffcients for the characteristics and time dummy variables in the hedonic case,

$$\boldsymbol{\beta}_{TD} = \begin{bmatrix} \alpha & \beta_1 & \dots & \beta_K & \tau_1 & \tau_2 & \dots & \tau_T \end{bmatrix}^T,$$

and equivalently the time dummy variable coefficients in the repeat sales case,

$$\boldsymbol{\beta}_{RS} = \begin{bmatrix} \delta_0 & \delta_1 & \dots & \delta_T \end{bmatrix}^T.$$

Note that p is the number of parameters in $\boldsymbol{\beta}$ and is different for the two models. We have that $p = K + T + 1$ for the hedonic time dummy model, and for the repeat sales model we have that $p = T + 1$.

Next, we introduce the matrix of the covariates,

$$\boldsymbol{x} = \begin{bmatrix} x_{11} & x_{21} & x_{p1} \\ x_{21} & x_{22} & x_{p2} \\ \vdots & \vdots & \vdots \\ x_{1N} & x_{2N} & x_{pN} \end{bmatrix},$$

and the random error vector

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}.$$

Again, the values of the matrix of covariates $\boldsymbol{x}$ and the values of the random error vector, $\boldsymbol{\epsilon}$ are different for the two models. However, the method

of computing the estimates for the parameter vector, $\hat{\boldsymbol{\beta}}$, for the two models is now easily demonstrated since we have introduced generalized notations.

We have that the models can be expressed in vector form as,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & x_{p1} \\ x_{21} & x_{22} & x_{p2} \\ \vdots & \vdots & \vdots \\ x_{1N} & x_{2N} & x_{pN} \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \tag{13}$$

Since we do not know the true parameter values in $\boldsymbol{\beta}$, we need to estimate (13), which yields,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & x_{p1} \\ x_{21} & x_{22} & x_{p2} \\ \vdots & \vdots & \vdots \\ x_{1N} & x_{2N} & x_{pN} \end{bmatrix} * \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}, \tag{14}$$

where the last term is the residual vector,

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}.$$

The residual vector, $\boldsymbol{e}$, is defined as the observed values in the response variable vector, $\boldsymbol{y}$, minus the fitted values from the regression, $\hat{\boldsymbol{y}} = \boldsymbol{x}\hat{\boldsymbol{\beta}}$, which is intuitive by looking at (14).

The parameter vector, $\boldsymbol{\beta}$, is estimated by minimizing the sum of the squares of the differences between observations and fitted values, or equivalently, minimizing the sum of the squared residuals,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) = \arg\min_{\beta}\boldsymbol{e}^T\boldsymbol{e}. \tag{15}$$

It is shown in Notes on Econometrics (2019, p.21) that the unique solution to (15) is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{y}, \tag{16}$$

which is called the *Ordinary Least Squares estimator*. The estimates produced by this estimator is required for the construction of indices in both methods.

For the estimates of the regression to be accurate, the following assumptions of the OLS regression must hold (OLS is short for Ordinary Least Squares),

1. *Linearity* - assuming that the model is linear in the parameters, β_k.

2. *Exogenity* - assuming that the errors are uncorrelated with the covariates, $Cov(\boldsymbol{x}, \boldsymbol{\epsilon}) = 0$.

3. *No autocorrelation* - assuming that the errors are uncorrelated with each other, meaning mathematically that $Cov(\epsilon_i, \epsilon_j | \boldsymbol{x}_{.i}, \boldsymbol{x}_{.j}) = 0$ for all $i \neq j$.

4. *No multicollinearity* - no covariates should be heavily correlated since they are supposed to be independent, $Cov(\boldsymbol{\beta}) = \sigma^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$.

5. *Normality and homoscedasticity* of the errors - assuming that the error terms are $\boldsymbol{\epsilon}|\boldsymbol{x} \sim N(0, \sigma^2\boldsymbol{I})$ with constant variance σ^2.

According to the Gauss-Markov theorem, when these assumptions hold, the OLS estimator $\hat{\boldsymbol{\beta}}$ is the *Best Linear Unbiased Estimator*, often shortened as BLUE (Rao, 1973, p. 281). For the OLS estimator to be unbiased entails that the expected values of the estimates are the true parameter values,

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

It is deemed the best linear estimator since it has the minimum variance of all linear unbiased estimators (Notes in Econometrics, 2019, p. 36). Furthermore, the OLS estimator is *consistent*, meaning that $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$ as $n \to \infty$. The proofs of these properties can be found in Notes in Econometrics (2019, p. 37)

However, oftentimes these assumptions are violated to some degree, then another regression method could be considered.

2.3.2 Weighted Least Squares

One should turn to the *weigthed least squares* regression method when dealing with a *heteroscedastic* data set (James et al, 2017, p. 96). This is applied when the assumption of homoscedasticity is violated and the error terms $\boldsymbol{\epsilon}$ do not have constant variance, instead the covariance matrix of the errors can be expressed as

$$Var(\boldsymbol{\epsilon}) = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \sigma_N \end{pmatrix}.$$

To account for this heteroscedacity, we introduce the *weight matrix*,

$$\boldsymbol{W} = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & w_N \end{pmatrix}.$$

The elements in the weight matrix is defined as $w_n = \frac{1}{\sigma_n^2}$. This way, when we multiply the weights with the squared residuals, we have given less weight to observations with residuals that have high variance and more weight to observations with low variance residuals,

$$\boldsymbol{\beta}_{wls} = \arg\min_{\beta} \sum_{i=1}^{N} w_i e_i^2. \tag{17}$$

It is shown in Notes in Econometrics (2019, p.89) that the solution to (17) is given by,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{wls}} = (\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{x})^{-1}\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{y}. \tag{18}$$

The intuition behind the weighted least squares is to transform the data to have homoscedasticity, and then run an ordinary least squares regression on the transformed data. In the repeat sales case, observations that have an abnormally long or short time span between the two sales would be given less weight in the regression due to depreciation and appreciation of the condominium. Mathematically this is done by following the procedure of Case and Shiller (1987, p. 15),

1. Run an Ordinary Least Square regression on the repeat sales model (9) and save the residuals from the regression. The residuals is the observed values minus the fitted values, $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{x}\hat{\boldsymbol{\beta}}$.

2. Run a new Ordinary Least Square regression of the squared residuals on a constant and the time span between the sales.

3. Now divide each observation in (9) with the square root of the fitted values from the regression in step 2. Then run another Ordinary Least Squares regression on this modified data.

This procedure will yield a weighted least squares estimator that accounts for heteroscedasticity and is the default weighting for the repeat sales method.

For the hedonic time dummy method however, there is no default way of weighting. Diewert (2005, p. 563) proposed using the expenditure share of the condominiums as weights. This means that each period t gets a weight representative of how large the share of the contract prices is sold during that period,

$$w_t = \frac{\text{sum of sales in period t}}{\text{sum of all sales}}.$$

However, there are many different ways of choosing weights for the time dummy hedonic model and I will not cover them in this thesis.

Noteworthy is that in the case that the error term is homoscedastic before going through with the weighting procedure, the resulting WLS (abbreviation for Weighted Least Squares) estimator can introduce heteroscedasticity rather than removing it (Silver, 2016, p. 51).

2.3.3 Robust regression

We have now discussed how to handle heteroscedasticity using weighted least squares regression in favor of ordinary least squares. But these regression methods are both sensitive to *outliers*, which are observations with large residuals. In other words, outliers are observations far away from their fitted values. An example could be in the case where the realtor has managed to raise the bidding price a lot, so much that the fitted value of the condominium from the model undermines the elevated bidding price. The residual of such an observation would then potentially be considered an outlier if the difference between the inflated actual price and the predicted price of the model is large enough.

Outliers can be common in a sample of housing and these observations can lead to inaccurate estimates of the coefficients and in result skewed house indices (Eurostat, 2013, p. 51). However, the *robust regression method* accounts for these outliers and this will be the last method I will introduce. The theoretical framework behind robust regression is quite wide and advanced. Therefore, I will merely provide the basics and intuition behind it. The following formulation of the robust regression theory is heavily influenced by the works of Fox and Weisberg (2011) and I refer to their paper for further details.

There are multiple approaches to robust regression, the one that we will cover is the M-estimation approach and will be applied when creating the robust indices in section 3. Similarily to OLS, M-estimation minimizes a sum of a function of the residuals $H(e_i)$. In the OLS case, the function is simply, $H(e_i) = \sum \epsilon_i^2$. However, in the robust regression, it is boils down to estimating the following equations,

$$\sum_{i=1}^{n} w_i(y_i - \boldsymbol{x_i^T}\hat{\boldsymbol{\beta}})\boldsymbol{x_i^T} = \mathbf{0}. \tag{19}$$

Here is $\boldsymbol{x_i^T}$ the transpose of the ith row in the explanatory variable matrix, whereas w_i is the weight function for observation i. The problem here is that the weights needed to solve the equation system is dependent on the residuals and vice versa. Therefore, an iterative solution method is implemented, called *iteratively reweighted least-squares*, abbreviated to IRLS. The procedure of this method is as follows,

1. Estimate initial values for $\hat{\boldsymbol{\beta}}^{(0)}$, using some sort of regression method, for example regular OLS estimation.

2. For each step, t, in the iteration process, calculate the residuals $e_i^{(t-1)}$ and the corresponding weights that are dependent on those residuals, $w_i^{(i-1)} = w(e^{(t-1)})$.

3. Then calculate new weighted least squares estimates for $\hat{\boldsymbol{\beta}}_{\boldsymbol{rob}}{}^{(t)}$ according to

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{rob}}{}^{(t)} = (\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{W}^{(t-1)}\boldsymbol{x})^{-1}\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{W}^{(t-1)}\boldsymbol{y}.$$

The last two steps are then iterated until the parameter estimates stabilize and converge (Fox and Weisberg, 2011). The weight function that is used in section 3 is called the *bisquare* weight function and is defined as,

$$w_B(e) = \begin{cases} (1 - (\frac{e}{k})^2)^2, & \text{for } |e| \leq k \\ 0, & \text{for } |e| > k. \end{cases} \tag{20}$$

In (20) the number k is called the tuning constant. This can be thought of as a threshold, the smaller the tuning constant is, the more is the estimates unaffected by outliers. However, taking a too small of a k will lead to decreased efficiency of the model, since many of the observations will then have weight equal to zero. Usually, for the bisquare weight function, the tuning constant is set to $k = 4.685\hat{\sigma}_\epsilon$, where $\hat{\sigma}_\epsilon$ is the estimated standard deviation of the errors (Fox and Weisberg, 2011). Following this robust regression procedure, we have down-weighted observations with large residuals and removed observations outside of the tuning constant k and hence accounted for outliers.

3 Application of Theory

In this section, I will examine the hedonic time dummy model and the repeat sales model when running these three regression methods.

3.1 Data

Data used in this study was provided by Hans Flink at Mäklarstatistik AB. I received a dataset containing $50,258$ observations on contract transactions in Stockholm municipality during the period from January 2014 to December 2018. The variables included in the data set are stated in Table 1 below.

Numerical Variables	Categorical Variables
Contract price	Congregation
Living area	District name
Price per sqm	Municipality
Building age	Formatted address
Monthly fee	Apartment floor
Contract date	Building storeys
Longitude	Number of rooms
Latitude	New production
	Balcony
	Elevator

Table 1: Variables in the data set

The categorical variables were treated as dummy variables with different number of levels. For example, the Number of rooms-variable had levels 1 to 9. Let us say that an observation had three rooms, the dummy variable denoting level Number of rooms = 3 would then be set to 1 and 0 for all of the other levels. In contrast, the numerical characteristics were treated as continuous variables.

The data contained quite a lot of observations that were gathered in a faulty manner. For example, the "building age" variable did occasionally contain multiple values within one cell, due to renovations and rebuilding. There were also values in the wrong format such as "Stone Age" or "1400-1500". Some values were clearly incorrect, such as apartments on floor 43533. These ambiguous formats would damage the hedonic time dummy regression and was therefore dropped for the hedonic case.

The result of the data wrangling and tidying resulted in a lowered the sample size of 38363 which was then used to create the hedonic index. This modification of the data was not necessary for the repeat sales modeling since the characteristics variables were not needed.

3.2 Hedonic Time Dummy Modeling

We will start by discussing the hedonic time dummy modeling. The characteristics in the model were chosen using a stepwise selection function in the software RStudio, called step(). The purpose of this function is to find the best model based on *Akaike's information criterion*, shortened as AIC. The formula for AIC is defined as

$$AIC = 2p - 2l(\boldsymbol{\theta_{ML}}), \tag{21}$$

where p denotes the number of parameters in the model and $l(\boldsymbol{\theta_{ML}})$ is the maximised log-likelihood function for the model. For further details on

why this is an appropriate metric to use when comparing models, see Held and Bové (2013, p. 224). The step function compares different sized models to each other and the model with the lowest AIC was chosen. Ultimately, the final hedonic model was

$$ln(p_n^t) = \alpha^t + \sum_{k=1}^{8} \beta_k z_{nk}^t + \sum_{\tau=1}^{60} \gamma^\tau D_n^\tau + \epsilon_n^t,$$

with the 8 characteristics found in Table 2 below.

Numerical Variables	Categorical Variables
Living area	District name
Building age	Apartment floor
Monthly fee	Number of rooms
	New production
	Elevator,

Table 2: Characteristics in the Hedonic Model

Furthermore, 60 time dummy variables representing each month from January 2014 to December 2018 was included in the model.

The district variable is a locational variable divided up in 18 areas around Stockholm. Each district along with how many observations were from there, is found in Table 8 in the Appendix.

Using this model we have controlled for the most important characteristics, except for materials used in the construction of the condominiums which was not included in the data (Eurostat et al, 2013, p. 25).

3.2.1 Analysis of the OLS Hedonic Model

The coefficient of determination for the OLS hedonic time dummy model was $R^2 = 0.904$. This value is a measure of how much of the variation in the data is explained by the model, it is defined as

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2}. \tag{22}$$

The sum in the numerator is the squared prediction errors of the model and the sum in the denominator is the sum of squared deviations of the observations from their mean (James et al, 2017, p. 70). Meaning that, using our hedonic time dummy model, we have explained more than 90% of the variation of the data.

Furthermore, all of the time dummy estimates were statistically significant on a 99% confidence level, except for two months; February and March of 2014. These months had p-values of 0.29 and 0.20 which are greater than 0.01 and are therefore not considered statistically significant.

If the p-value for a covariate β_i is smaller than 0.01, the null hypothesis that $H_0 : \beta_i = 0$ is rejected in favor of the alternative hypothesis that $H_a : \beta_i \neq 0$ on a 99% confidence level. This implies that there is a statistically significant relationship between the response variable and the covariate. In contrast, if the p-value is greater than 0.01, we cannot say with 99% confidence that there is an association between the covariate and the response variable (James et al, 2013, p.67).

The estimates, standard errors and, p-values are found in Table 9, Table 10, and Table 11 in the appendix. Additionally, the number of observations for each month is also illustrated in Figure 10 in the appendix.

3.2.2 Residual plots for the OLS Hedonic Model

Now, we will check how well the hedonic model fulfills the OLS model assumptions of section 3.2.1, by examining the residual plots.

The linearity assumption of the OLS hedonic model can be evaluated in the Residual vs Fitted plot of Figure 1. Here we want to see a straight red line at 0 and no clear patterns (Colonescu, 2016). When many observations are close to the zero line, this implies that the observations have small residuals and the fitted values are close to the observed values. This would indicate a linear relationship between the dependent variable and the covariates since we have fitted a multilinear model. Note, in the Residual vs Fitted plot of Figure 1 below, that the linearity assumption is approximately fulfilled, with a slight deviation downward for larger fitted values. This reveals that our model underestimates the higher observed values, since the residuals are more negative for the higher fitted values.

From the Normal Q-Q plot in Figure 1 we will evaluate the assumption of normality of the residuals. It is called the Normal Q-Q plot since it compares the theoretical quantiles of the normal distribution (x-axis) to the empirical quantiles of the *standardized residuals* (y-axis). The standardized residuals are defined as the residuals divided by the sample standard deviation (Martin et al, 2017, p. 135). Ideally, the Normal Q-Q plot would illustrate a linear line, which would imply that the standardized residuals are normally distributed. However, we see in the Normal Q-Q plot of Figure 1, that this assumption is somewhat violated due to its' tail deviating from the straight line to the left.

The Scale-Location plot gives us information on how well the model holds the homoscedasticity assumption. On the y-axis is the square root of the standardized residuals and on the x-axis is the fitted values. Ideally, the plot would illustrate a horizontal line with evenly spread points, since this implies that the residuals have constant variance. When we look at the Scale-Location plot of Figure 1, it is noticeable that the assumption of homoscedasticity is somewhat violated since the points are not evenly spread out for higher fitted values.

In the fourth plot of Figure 1, Residuals vs Leverage, we can examine if the model is subject to any outliers or high leverage points. Outliers are observations that have large residuals, in this case meaning that the price of the observation is extreme in comparison to the predicted price of the condominium. A high leverage point is an observation with extreme values in the explanatory variables. In this plot, the dotted red line is the *Cook's distance line*, this is a metric that combines the leverage and outlier influence on the model. The Cook's distance is defined as

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y_j} - \hat{y}_{j(i)})^2}{pMSE}, \tag{23}$$

where $\hat{y_j}$ is the fitted value of observation j, whereas $\hat{y}_{j(i)}$ is the fitted value of observation j when observation i is excluded (Neter et al, 1996, p. 402). Meaning that, in the nominator of (22), we have the squared differences between the fitted values when excluding an observation i. The bigger this difference is for an observation i, the more influence will that observation have on the regression. In the denominator, p is the number of covariates and MSE is the mean squared error of the regression.

In the Residuals vs Leverage plot, observations outside of the red dotted line have high Cook's distance values and therefore high influence on the model. Note, in the residuals vs leverage plot of Figure 1, that there is one observation outside of the dotted line.

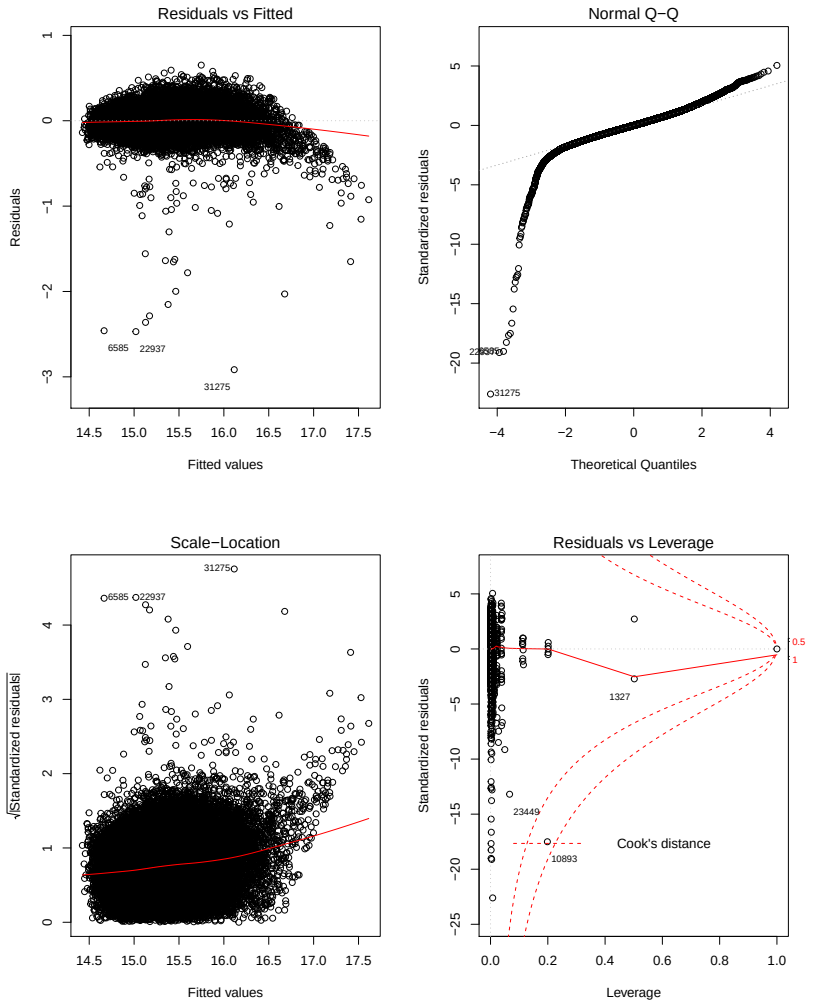

Figure 1: Residuals OLS Hedonic Model

3.2.3 Analysis of the WLS and Robust Hedonic Model

There is no default way of choosing weights for the WLS (short for Weighted Least Squares) hedonic time dummy model, hence I experimented and used the expenditure shares as weights, as proposed by Diewert (2005, p. 563).

Since we have concluded from these residual plots that the assumptions of the OLS hedonic model are violated, we continue to examine the residual plots for the WLS hedonic model. The residual plots of this regression is shown in Figure 2 below. We can see that the plots are very similar to the plots in Figure 1. There still seems to be heteroscedasticity in our data, judging from the Scale-Location plot of Figure 2. By examining the Residuals vs Leverage plot of Figure 2, there seems to be an improvement as

the points are further away from the Cook's distance lines, this insinuates that the weighting has improved the handling of points with high leverage and large residuals. However, despite the weighting procedure, there is still one observation outside of the Cook's distance line.

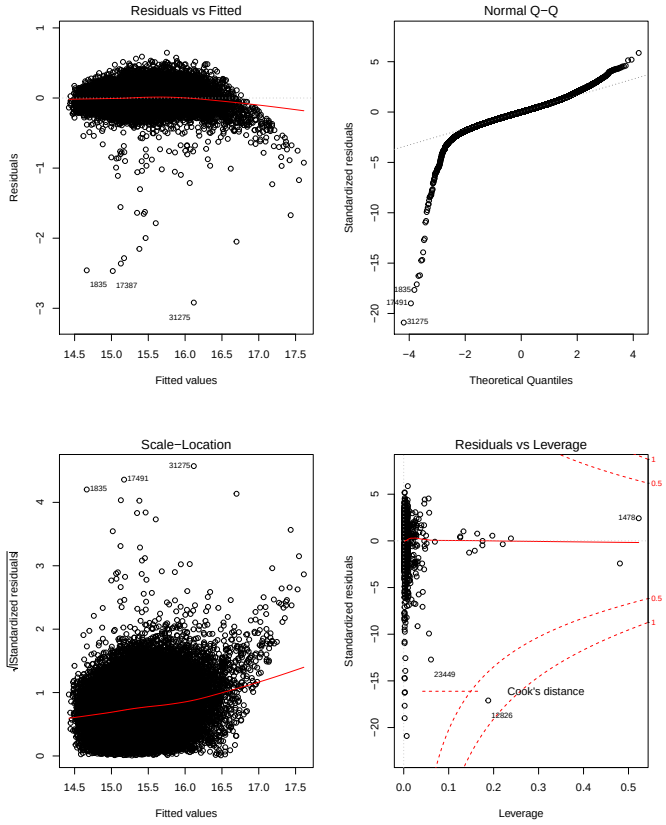

Figure 2: Residuals WLS Hedonic Model

In the robust model, observations with large residuals have been given weight 0 in accordance with the theory in section 2.3.3. This has resulted in estimates for the robust regression with smaller standard errors, some of the time dummy variable standard errors are illustrated in Table 3.

Variables	OLS Standard Error	WLS Standard Error	Robust Standard Error
2014-10	0.0076	0.0090	0.0066
2014-11	0.0079	0.0095	0.0067
2014-12	0.0091	0.0129	0.0078
2014-2	0.0082	0.0102	0.0071
2014-3	0.0079	0.0096	0.0068
2014-4	0.0080	0.0099	0.0069

Table 3: HTD: Comparison of standard errors for a few estimates

3.2.4 Index construction for the Hedonic Model

Following the theory provided in section 2.1, three hedonic indices were created using the hpiR package in R created by Andy Krause. The results are found in Figure 3 below.

Figure 3: Hedonic Time Dummy Indices

From Figure 3, we notice that the index had a steady upward trend from January 2014 to July 2017 where it had its peak. Followed by a decreasing trend from July 2017 towards the end of 2017, which resulted in a total drop of 15%. We can see that the all of the hedonic indices are very similar, meaning that all three regression methods yielded very similar estimates for the time dummy coefficients.

3.3 Repeat Sales Modeling

In unison with the theory of the repeat sales equation in section 2.2, the repeat sales model was set up as previously with the same notations,

$$ln(p_n^t) - ln(p_n^s) = \sum_{t=0}^{60} \delta^t D_n^t + \mu_n^t, \text{ for } n \in \{1, 2, ..., N\}. \tag{24}$$

We have that t ranges from 0 to 60 since we have observations from January 2014 to December 2018.

At first, I ran the regression on the data set containing all of the repeat sales data, which was $5,420$ observations. This data set included observations with short holding periods and was therefore subject to house flips and appreciation issues. The summary output for this model had several estimates that were not statistically significant on a 99% confidence level. Furthermore, the coefficient of the determination of this model was very low, merely $R^2 = 0.297$, meaning that the model did not explain the variation in this data set very well.

To improve my repeat sales model I followed the suggestions of Clapp and Giacotto and filtered my data for only observations with a minimum holding period of 12 months to account for the house flipping issue. Meaning that the period between the sale and the resale was at least one year for all observations. This resulted in a data set containing 2859 observations. The summary output for this repeat sales model had a higher coefficient of the determination at $R^2 = 0.5259$ and the only estimates that were not statistically significant on a 95% confidence level was February, April, and May of 2014. The estimates, standard deviation of the estimates, and the p-values can be found in Table 12 and Table 13 in the appendix.

3.3.1 Analysis of the Repeat Sales Model

Similarly to the hedonic modeling, I evaluated the residual plots for the repeat sales model, in Figure 4. In the Residuals vs Fitted plot of Figure 4, we see a horizontal line with randomly distributed points above and below, which tells us that the model is fulfilling the OLS assumption of linearity. The Scale-Location plot also looks as desired, suggesting homoscedasticity in the model. The Residuals vs Leverage plot looks appropriate as well,

with observations inside the Cook's distance line. However, the Normal Q-Q-plot has tail deviations, suggesting that the residuals are not normally distributed.

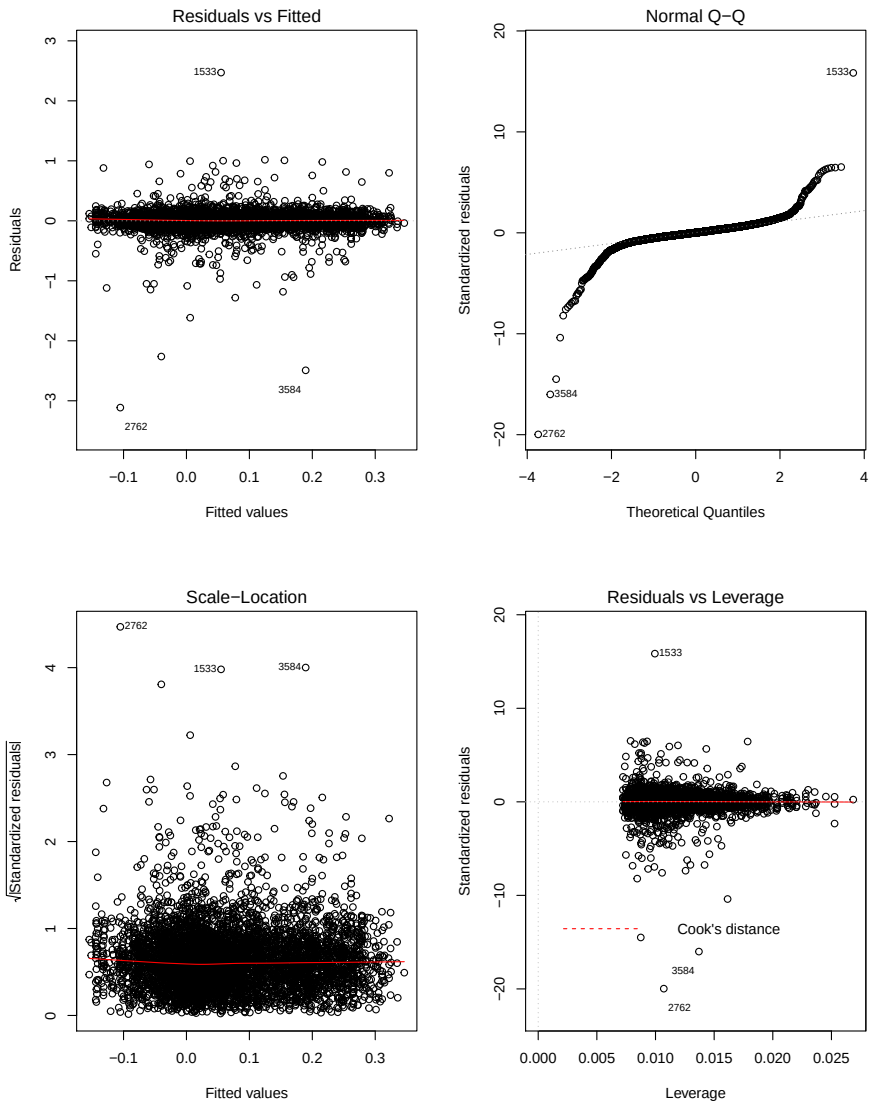

Figure 4: Residual plots OLS Repeat Sales Model

Although the residual plots for the OLS regression look desirable, I included the residual plots for the WLS regression as well in Figure 5, to see whether or not the weighting harmed the residual plots. By analyzing Figure 5, we can see that the plots remain similar to the OLS case.

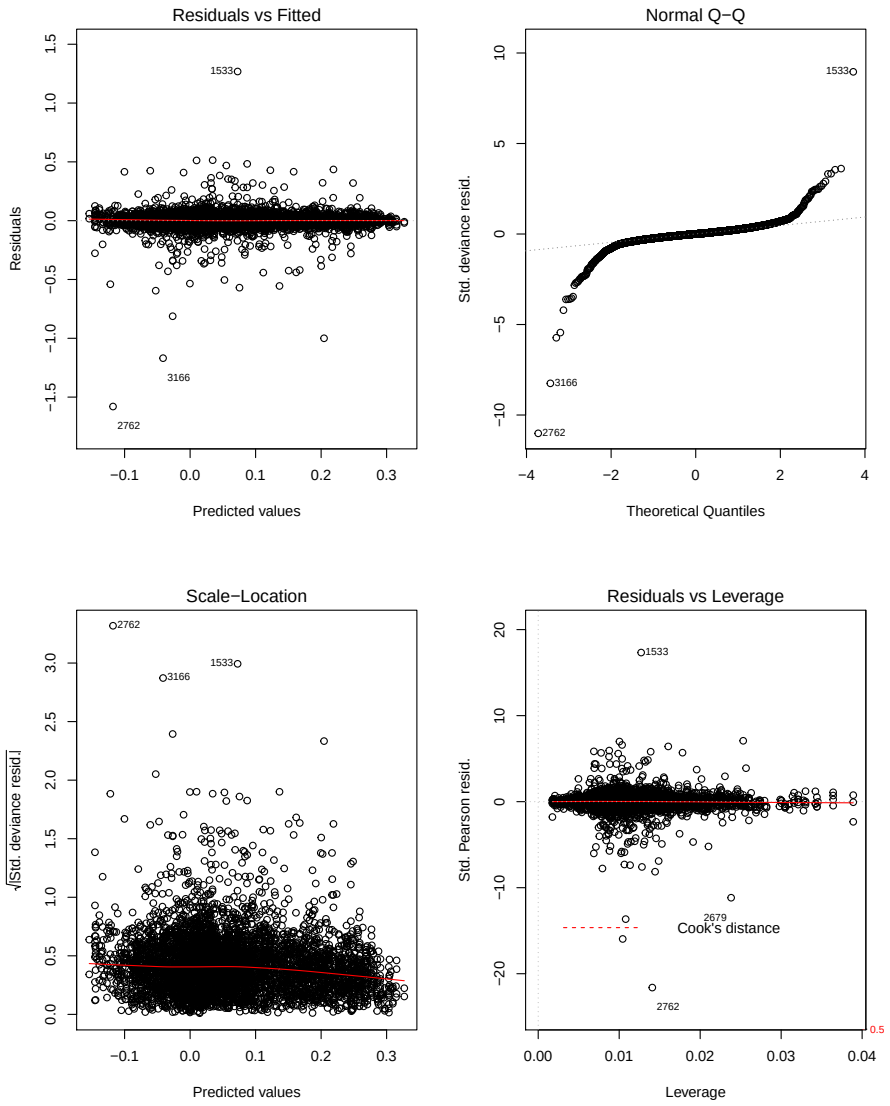

Figure 5: Residual plots WLS Repeat Sales Model

Again, the standard errors of the estimates in the robust regression model was lower, due to the exclusion of outliers, as illustrated in Table 4.

Period Variable	OLS Standard Error	Robust Standard Error
1	0.022	0.019
2	0.024	0.020
3	0.028	0.022
4	0.027	0.021
5	0.026	0.020
6	0.026	0.018

Table 4: RS: Comparison of standard errors for a few estimates

3.3.2 Index construction for the Repeat Sales Model

Once again, the hpiR package by Andy Krause was used to produce three repeat sales indices, one for each regression method. The results are found in Figure 6.

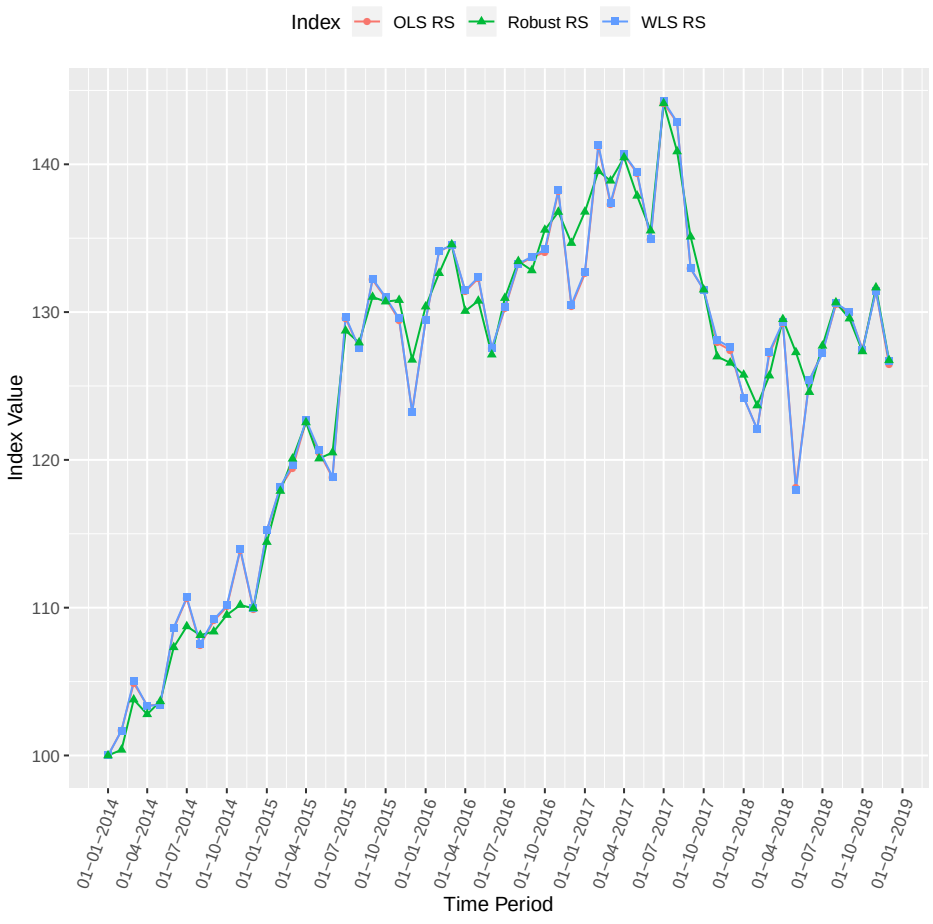

Figure 6: Repeat Sales Indices

By looking at Figure 6, we can see that the Ordinary Least Squares and Weighted Least Squares regressions yield almost identical repeat sales indices and that the robust regression index follows the same movement trends but less drastically, similarly to the hedonic case.

3.4 Comparison of indices

In Figure 7, we have compared the OLS repeat sales index with the OLS hedonic time dummy index.

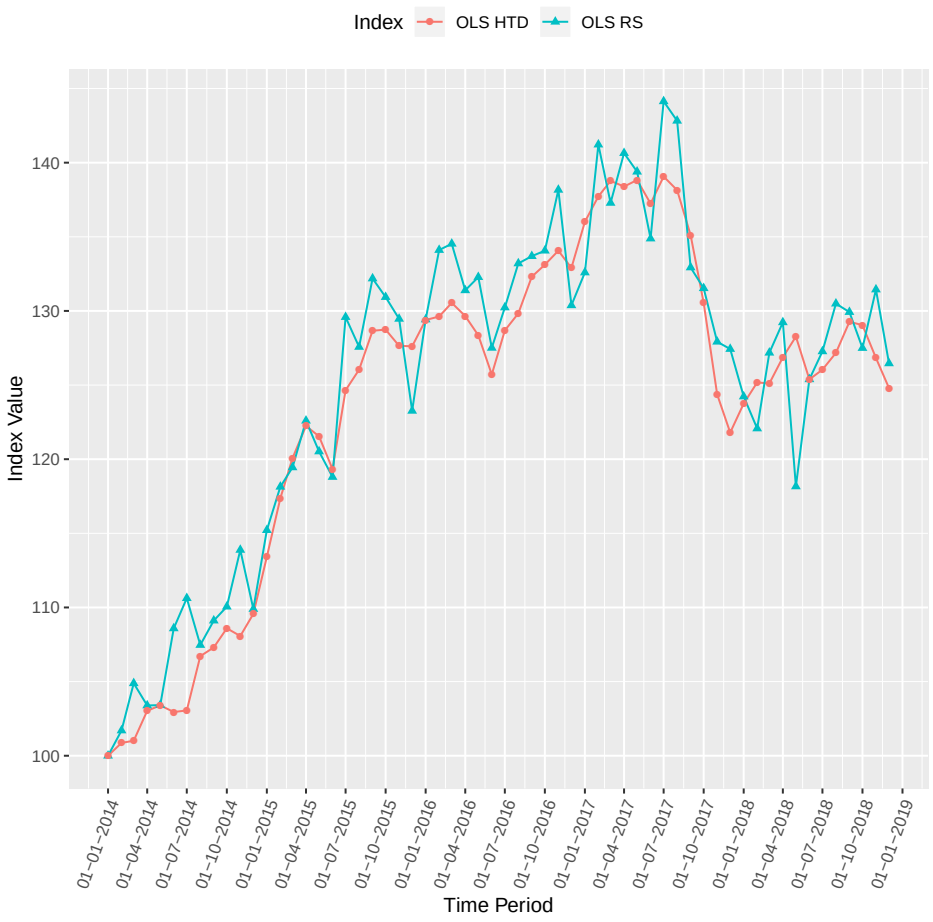

Figure 7: Repeat Sales vs Hedonic Time Dummy Index

We note from Figure 7 that the indices tend to follow the same trends, but that the repeat sales index is more volatile. Comparisons of the WLS and robust indices can be found in Figure 11 and Figure 12 in the appendix. By examining Figure 12, we note that the robust indices for the two models are the most similar.

4 Validation of the Indices

When considering the validity of an index, Andy Krause proposes three metrics to compare different indices; accuracy, volatility, and revision (Krause, 2019). In this section I will introduce these three validation metrics and use them to compare the indices I have created. I finish this section by comparing the indices with a published index for condominiums in Stockholm, called the HOXFLATSTO.

4.1 Accuracy

In the hpiR package, the accuracy of an index is calculated by looking at the median of errors between repeat sales pairs. Which is defined in this case as the predicted price of the second sale adjusted from the first sale price using the index, minus the actual price of the condominium at the second sale,

$$E_i = ln(p_{i,predicted}) - ln(p_{i,actual}). \tag{25}$$

The smaller the errors are, the more accurate the index is deemed to be (Krause, 2019). In Table 5 I have gathered the results of the accuracy tests for the six indices. The accuracy is calculated using a K-fold test. This entails that some percentage of the data is removed, in this case I used 10% as recommended by the author of the package. After the removal of this part of the data, the index is then estimated again in the same method as before (hedonic for the hedonic indices and repeat sales for the repeat sales indices). This newly created index on the smaller data sample is then used to predict the sample that was removed. The errors between these predicted prices and the actual prices are then calculated and the median of the absolute errors, MAE, is shown in Table 5.

	K-fold: MAE
RS: OLS	0.05934568
RS: WLS	0.05958107
RS: Robust	0.05688765
HTD: OLS	0.05776318
HTD: WLS	0.05784941
HTD: Robust	0.05639458

Table 5: Test of Accuracy, median absolute error

From Table 5 we can conclude that all of the indices have similar accuracy, but that the hedonic time dummy indices are slightly more accurate than the repeat sales indices. Furthermore, it seems like the robust regression method produces the most accurate indices.

4.2 Volatility

The second metric is measuring the volatility of the index, which is done by calculating the standard deviation of the index changes in a rolling four months period,

$$Vol_{index} = sd(D_t, D_{t+1} + D_{t+2}), \tag{26}$$

where, $D_t = P_t - P_{t-1}$, is the changes in the index from period $t-1$ to t (Krause, 2019). Using this metric, consistent movement over a 4 month time span will yield low volatility, whereas more drastic and non monotonic changes will yield higher volatility. In Table 6 is the result from the volatility test gathered.

	Mean Volatility	Median Volatility
Repeat Sales: OLS	0.03397095	0.02969578
Repeat Sales: WLS	0.02695082	0.02535953
Repeat Sale: Robust	0.02147779	0.02029811
Hedonic: OLS	0.01347982	0.01120935
Hedonic: WLS	0.01345231	0.0112093
Hedonic: Robust	0.0123226	0.009522221

Table 6: Volatility of Indices

From Table 6 we conclude that the OLS and WLS indices in both methods have similar volatility values, whereas the values for the robust index are lower. This is not surprising judging from the previous graphs of the indices.

4.3 Revision

As we have previously discussed, both the hedonic time dummy and repeat sales indices suffer from the revision problem. Each time new data is added, the regression has to be re-run, which could alter already published indices. When an index is very stable however, this should not lead to any major revision. Therefore, revision is the third metric we will examine.

In the hpiR package, the function calcRevision() calculates the mean revision for each period. In this case, we have 60 periods with data, however, the function begins with calculating the index on merely the 12 first periods. Then one more period is added to the data and the index is re-calculated. The mean revision, for period $1-12$, is calculated as

$$R_{12} = \frac{\sum_{i=1}^{12}(P_{i,old} - P_{i,new})}{12},$$

where $\sum_{i=1}^{12} P_{i,old}$ is the 12 first indices from the regression based on only the 12 first periods, and $\sum_{i=1}^{12} P_{i,new}$ is the 12 new indices from the regression based on the 13 first periods. This scheme then continues, mean revisions for $1-13$, $1-14$ up to $1-60$ are calculated (Krause, 2019). The results of the mean of the mean revisions for the indices are found in Table 7 below.

Method	OLS	WLS	Robust
Hedonic Revision Means	0.00351	0.00351	0.00355
Repeat Sale Revision Means	0.0358	0.0407	0.04791

Table 7: Revision Means Comparison

Note that the robust indices are subject to more revision than the corresponding OLS and WLS indices. Additionally, the hedonic time dummy indices have lower revision means than the corresponding repeat sales indices.

4.4 Comparison to Nasdaq OMX Valueguard KTH Housing Index

The *Nasdaq OMX Valueguard-KTH Housing Index Flats Stockholm*, also known as HOXFLATSTO, is often referenced in Swedish media when statements concerning the current situation in the housing market in Stockholm is made. It is constructed by Valueguard Index Sweden AB and Svensk Maklarstatistik AB, with the aim to provide a reliable benchmark for the Stockholm condominium market (HOXTM, 2010). Since the HOXFLATSTO index is widely recognized, we will compare this index with the 6 indices constructed in this thesis. The comparison will serve as the final validation metric.

The HOXFLATSTO index is based on a hedonic price model (HOXTM, 2010) and should therefore be similar to my hedonic time dummy index. The published HOXFLATSTO index points were gathered from the Nasdaq OMX Nordic website (HOXFLATSTO, 2020). In Figure 8 below we can see the comparison between the HOXFLATSTO index and my hedonic time dummy indices. Note that the indices seem to follow the same trends, but that the HOXFLATSTO index seem to be higher than the hedonic time dummy indices that we have constructed.

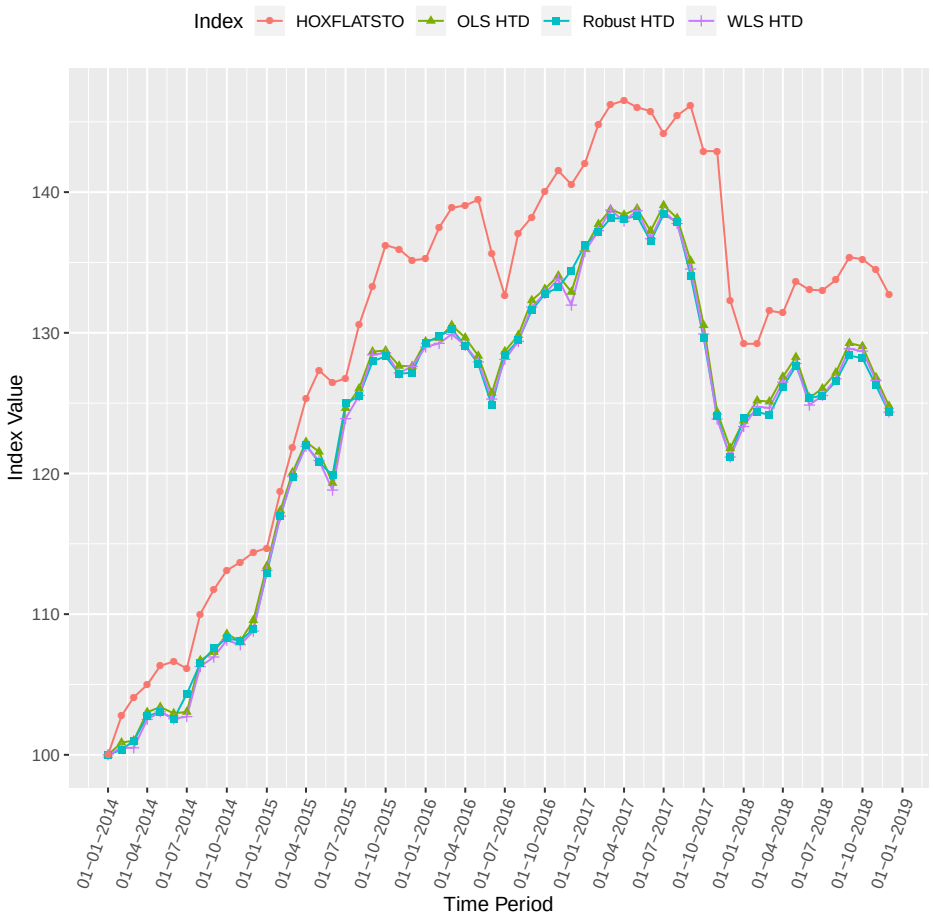

Figure 8: Comparison HOXFLATSWE and Hedonic TD index

When using a price index for evaluating the market status, it should represent the trends and changes in the pricing market over time. However, the actual value of different indices when compared should not necessarily be expected to be equal, as long as they follow the same trends, which in Figure 8, is clear that they do.

In Figure 9, we can see a comparison between the repeat sales indices and the HOXFLATSTO index. Again, the same trends are followed by all indices, however, the repeat sales indices are more volatile than the HOXFLATSTO index.

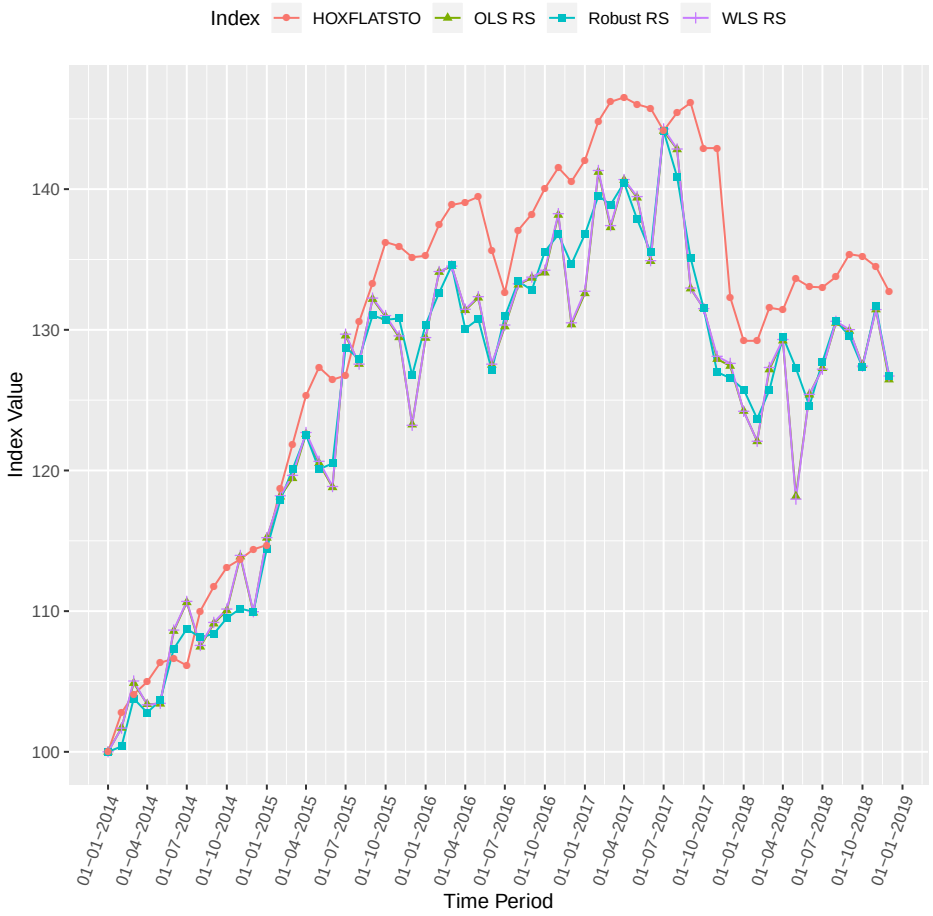

Figure 9: Comparison HOXFLATSWE and Repeat Sales index

To further compare the trends of my indices with the ones captured by HOXFLATSTO, the correlation coefficients between the robust indices and the HOXFLATSTO index was computed. The correlation coefficient between two indices p_1 and p_2 is calculated by,

$$\rho_{p_1,p_2} = \frac{Cov(p_1, p_2)}{\sigma_{p_1}\sigma_{p_2}}.$$

Where $Cov(p_1, p_2)$ is the covariance between the indices, whereas the σ_{p_1} and σ_{p_2} are the standard deviations for index p_1 and p_2 respectively (Sundberg, 2016, p. 89). Letting $p_1 = P_{TD}$, where P_{TD} is the robust hedonic time dummy index and p_2 be the HOXFLATSTO index. The correlation coefficient was computed to be 0.98, which indicates a strong positive correlation between these indices. When the robust hedonic time dummy index is high, the HOXFLATSTO index tends to be high as well, and vice versa.

Similarly, the correlation coefficient between the HOXFLATSTO index and the robust repeat sales index was computed and the result was equal to 0.96, which also indicates a strong positive correlation between the indices.

5 Discussion

To conclude, in this thesis we have constructed six different indices based on two different price models, the repeat sales model and the hedonic time dummy model. We have discussed the structure, the index construction process, and the drawbacks of these models. Furthermore, we have examined how well the models fit the assumptions of three different regression methods for our given data, the Ordinary Least Squares, Weighted Least Squares, and Robust regression.

5.1 Summary of Hedonic Time Dummy Indices

The residuals plots for the hedonic model in Figure 1 revealed that some of the Ordinary Least Squares assumptions were violated. In Notes in Econometrics (2019, p.70), it is stated that if the assumption of homoscedasticity is violated, the estimator, $\hat{\boldsymbol{\beta}}_{ols}$ remains unbiased. However, it may not remain BLUE, meaning that there can be another linear unbiased estimator with a lower variance. We tried to account for heteroscedasticity by implementing a weighted least squares regression. But as the residual plots in Figure 2 revealed, it did not resolve the issue. However, they may not have the smallest variances out of all estimators, but since we have a sufficiently large data set the issue of heteroscedasticity should not be exaggerated. The standard errors of the estimates are still low, see Table 9 and Table 10 in the appendix. Meaning that the estimates are precise, despite of the heteroscedasticity. The robust hedonic time dummy index had even lower standard errors of the estimates, suggesting that this regression method produced the most precise estimates.

In Section 4, we evaluated the hedonic index based on three validation metrics; accuracy, volatility, and revision. In the accuracy and volatility tests, the robust hedonic time dummy index performed best with the smallest errors. In contrast, the robust index performed slightly worse than its' competitors in the revision test.

Overall, in Figure 3, we can see that all three regression methods yield similar indices. Furthermore, in Figure 8, we see that the indices follow the same movement patterns as the HOXFLATSTO index, which is additional confirmation that our indices are precise and accurate. The difference in index value of the HOXFLATSTO index and the hedonic time dummy indices constructed in this thesis may be explained by different hedonic specifications for the models. The hedonic specification of the model for the HOXFLATSTO index is to this thesis unknown, perhaps they have

used another specification. For example, other covariates, such as building materials and coordinates, could have been used in the construction of the HOXFLATSTO index. Perhaps, they transformed some covariates, for example, squaring or taking the natural logarithm of numerical variables. Using another specification would lead to other estimates and hence a different value of the index. The creators of HOXFLATSTO may have also used a different sized sample and cleaned the data in another way. Potentially, this could have solved the issue of large negative residuals for higher fitted values.

5.2 Summary of Repeat Sales Indices

In Figure 4 of Section 3.3.1, the residuals of the OLS repeat sales model looked desirable and the assumptions seemed to be approximately fulfilled. The weighting procedure as proposed by Case and Shiller (1987, p. 15) of section 2.3.2 did not seem to introduce any heteroscedasticity, based on analysis of Figure 5. We saw in Figure 6 that the OLS and WLS repeat sales indices were almost identical, whereas the robust repeat sales index was less volatile. As in the hedonic case, the robust repeat sales index outperformed the OLS and WLS repeat sales indices in the accuracy and volatility test, but lost in the revision test.

It is clear that the HOXFLATSWE and the repeat sales indices differ and that the repeat sales indices are more volatile, from Figure 9. Nevertheless, they had a strong positive correlation which is positive validation for the repeat sales indices. However, these indices should not be expected to be the same, since they are constructed using different price models.

5.3 Potential improvements

The thesis could be improved by further analyzing the specification of the hedonic model. Perhaps it would be appropriate to transform some of the covariates to better fulfill the OLS assumptions. An example could be to include squares of numerical variables and check how that affects the residual plots.

Furthermore, additional analysis of the assumptions of each regression method could have been made, such as tests for autocorrelation and multicollinearity. Instead of merely analyzing residual plots. This was left out since the consequences of these violations do not affect that the estimates are unbiased (Notes in Econometrics, 2019, pp. 70-72), and the main purpose of this thesis was to construct trustworthy price indices which relies on unbiased estimates.

There could also be a more detailed look at outliers and leverage points, and experimenting with removing these points to see how that would affect the results. However, in this thesis I chose to use the robust regression

method as a way of handling problematic observations.

6 Appendix

Table 8: Number of observations from each district

District name	Number of observations in district
Essinge	1282
Högalid	3399
Kungsholmen	2471
Nacka	52
Skarpnäck	5
Adolf Fredrik	1115
Domkyrksdistrikt	217
Engelbrekt	2429
Gustav Vasa	1698
Hedvid Eleonora	933
Katarina	3379
Maria Magdalena	1668
Oscar	3900
Sankt Göran	5686
Sankt Johannes	1594
Sankt Matteus	4458
Sofia	4068
Sundbyberg	10

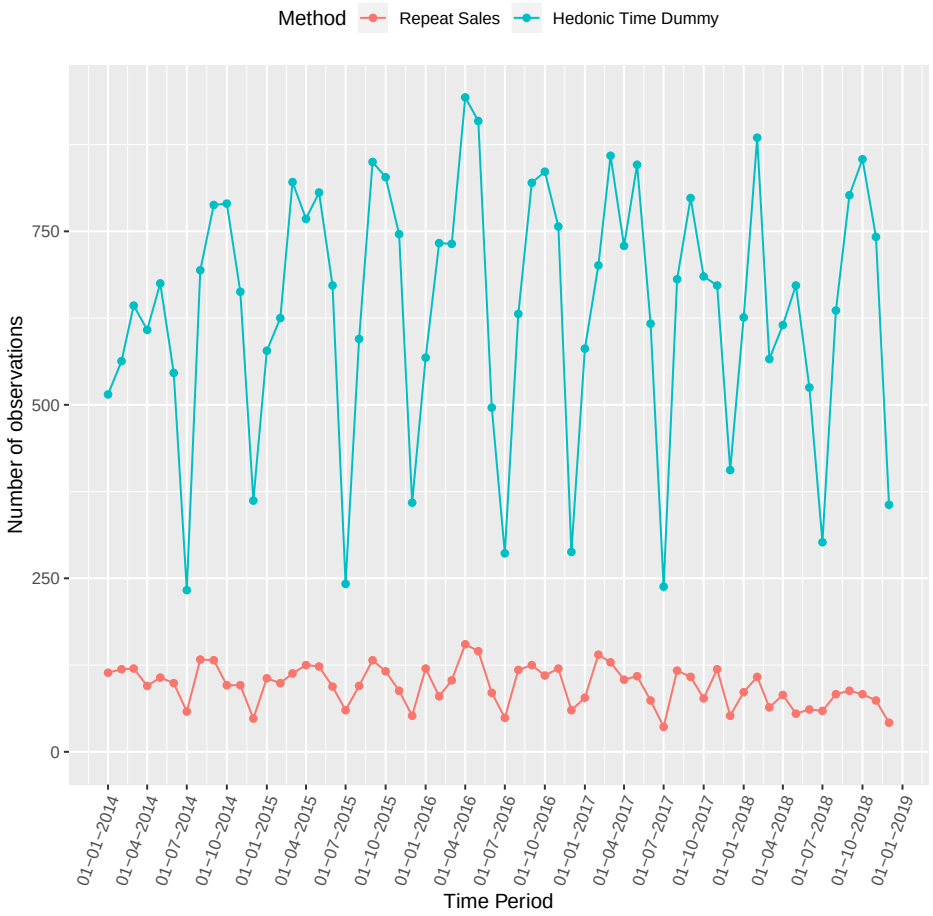

Figure 10: Number of observations for each month

Table 9: Hedonic Model: Time Dummy Variables (2014-2015)

Variable	Estimate	Standard Error	P-value
2014-10	0.08	0.01	0.00
2014-11	0.08	0.01	0.00
2014-12	0.09	0.01	0.00
2014-2	0.01	0.01	0.29
2014-3	0.01	0.01	0.21
2014-4	0.03	0.01	0.00
2014-5	0.03	0.01	0.00
2014-6	0.03	0.01	0.00
2014-7	0.03	0.01	0.00
2014-8	0.06	0.01	0.00
2014-9	0.07	0.01	0.00
2015-1	0.13	0.01	0.00
2015-10	0.25	0.01	0.00
2015-11	0.24	0.01	0.00
2015-12	0.24	0.01	0.00
2015-2	0.16	0.01	0.00
2015-3	0.18	0.01	0.00
2015-4	0.20	0.01	0.00
2015-5	0.20	0.01	0.00
2015-6	0.18	0.01	0.00
2015-7	0.22	0.01	0.00
2015-8	0.23	0.01	0.00
2015-9	0.25	0.01	0.00

Table 10: Hedonic Model: Time Dummy Variables (2016-2018)

Variable	Estimate	Standard Error	P-value
2016-1	0.26	0.01	0.00
2016-10	0.29	0.01	0.00
2016-11	0.29	0.01	0.00
2016-12	0.28	0.01	0.00
2016-2	0.26	0.01	0.00
2016-3	0.27	0.01	0.00
2016-4	0.26	0.01	0.00
2016-5	0.25	0.01	0.00
2016-6	0.23	0.01	0.00
2016-7	0.25	0.01	0.00
2016-8	0.26	0.01	0.00
2016-9	0.28	0.01	0.00
2017-1	0.31	0.01	0.00
2017-10	0.27	0.01	0.00
2017-11	0.22	0.01	0.00
2017-12	0.20	0.01	0.00
2017-2	0.32	0.01	0.00
2017-3	0.33	0.01	0.00
2017-4	0.32	0.01	0.00
2017-5	0.33	0.01	0.00
2017-6	0.32	0.01	0.00
2017-7	0.33	0.01	0.00
2017-8	0.32	0.01	0.00
2017-9	0.30	0.01	0.00
2018-1	0.21	0.01	0.00
2018-10	0.25	0.01	0.00
2018-11	0.24	0.01	0.00
2018-12	0.22	0.01	0.00
2018-2	0.22	0.01	0.00
2018-3	0.22	0.01	0.00
2018-4	0.24	0.01	0.00
2018-5	0.25	0.01	0.00
2018-6	0.23	0.01	0.00
2018-7	0.23	0.01	0.00
2018-8	0.24	0.01	0.00
2018-9	0.26	0.01	0.00

Table 11: Hedonic Model Summary Output:Characteristics

Variable	Estimate	Standard Error	P-value
(Intercept)	14.07	0.01	0.00
Högalid	0.10	0.00	0.00
Kungsholmen	0.15	0.00	0.00
Nacka	-0.22	0.02	0.00
Skarpnäck	0.01	0.06	0.85
Adolf Fredrik	0.20	0.01	0.00
Domhyrkodistrikt	0.06	0.01	0.00
Engelbrekt	0.11	0.00	0.00
Gustav Vasa	0.19	0.00	0.00
Hedvig Eleonora	0.28	0.01	0.00
Katarina	0.11	0.00	0.00
Maria Magdalena	0.12	0.00	0.00
Oscar	0.17	0.00	0.00
Sankt Göran	0.08	0.00	0.00
Sankt Johannes	0.15	0.01	0.00
Sankt Matteus	0.17	0.00	0.00
Sofia	0.06	0.00	0.00
Sundbyberg	-0.41	0.04	0.00
Living area	0.01	0.00	0.00
Monthly fee	-0.00	0.00	0.00
Age of building	0.00	0.00	0.00
New productionc1	0.12	0.01	0.00
Apartment floor	0.01	0.00	0.00
Number of rooms 2	0.13	0.00	0.00
Number of rooms 3	0.19	0.00	0.00
Number of rooms 4	0.17	0.00	0.00
Number of rooms 5	0.06	0.01	0.00
Number of rooms 6	-0.21	0.01	0.00
Number of rooms 7	-0.39	0.03	0.00
Number of rooms 8	-1.16	0.09	0.00
Number of rooms 9	-2.71	0.13	0.00
Elevator 1	0.02	0.00	0.00

Table 12: Repeat Sales Model Summary Output, 1-30

Period	Estimate	Standard error	P-value
1	0.02	0.02	0.39
2	0.05	0.02	0.02
3	0.03	0.02	0.11
4	0.03	0.02	0.10
5	0.08	0.02	0.00
6	0.10	0.02	0.00
7	0.07	0.02	0.00
8	0.09	0.02	0.00
9	0.10	0.02	0.00
10	0.13	0.02	0.00
11	0.09	0.03	0.00
12	0.14	0.02	0.00
13	0.17	0.02	0.00
14	0.18	0.02	0.00
15	0.20	0.02	0.00
16	0.19	0.02	0.00
17	0.17	0.02	0.00
18	0.26	0.02	0.00
19	0.24	0.02	0.00
20	0.28	0.02	0.00
21	0.27	0.02	0.00
22	0.26	0.02	0.00
23	0.21	0.02	0.00
24	0.26	0.02	0.00
25	0.29	0.02	0.00
26	0.30	0.02	0.00
27	0.27	0.02	0.00
28	0.28	0.02	0.00
29	0.24	0.02	0.00
30	0.26	0.03	0.00

Table 13: Repeat Sales Model Summary Output, 31-59

Period	Estimate	Standard error	P-value
31	0.29	0.02	0.00
32	0.29	0.02	0.00
33	0.29	0.02	0.00
34	0.32	0.02	0.00
35	0.27	0.02	0.00
36	0.28	0.02	0.00
37	0.35	0.02	0.00
38	0.32	0.02	0.00
39	0.34	0.02	0.00
40	0.33	0.02	0.00
41	0.30	0.02	0.00
42	0.37	0.03	0.00
43	0.36	0.02	0.00
44	0.28	0.02	0.00
45	0.27	0.02	0.00
46	0.25	0.02	0.00
47	0.24	0.03	0.00
48	0.22	0.02	0.00
49	0.20	0.02	0.00
50	0.24	0.02	0.00
51	0.26	0.02	0.00
52	0.17	0.02	0.00
53	0.23	0.02	0.00
54	0.24	0.02	0.00
55	0.27	0.02	0.00
56	0.26	0.02	0.00
57	0.24	0.02	0.00
58	0.27	0.02	0.00
59	0.23	0.03	0.00

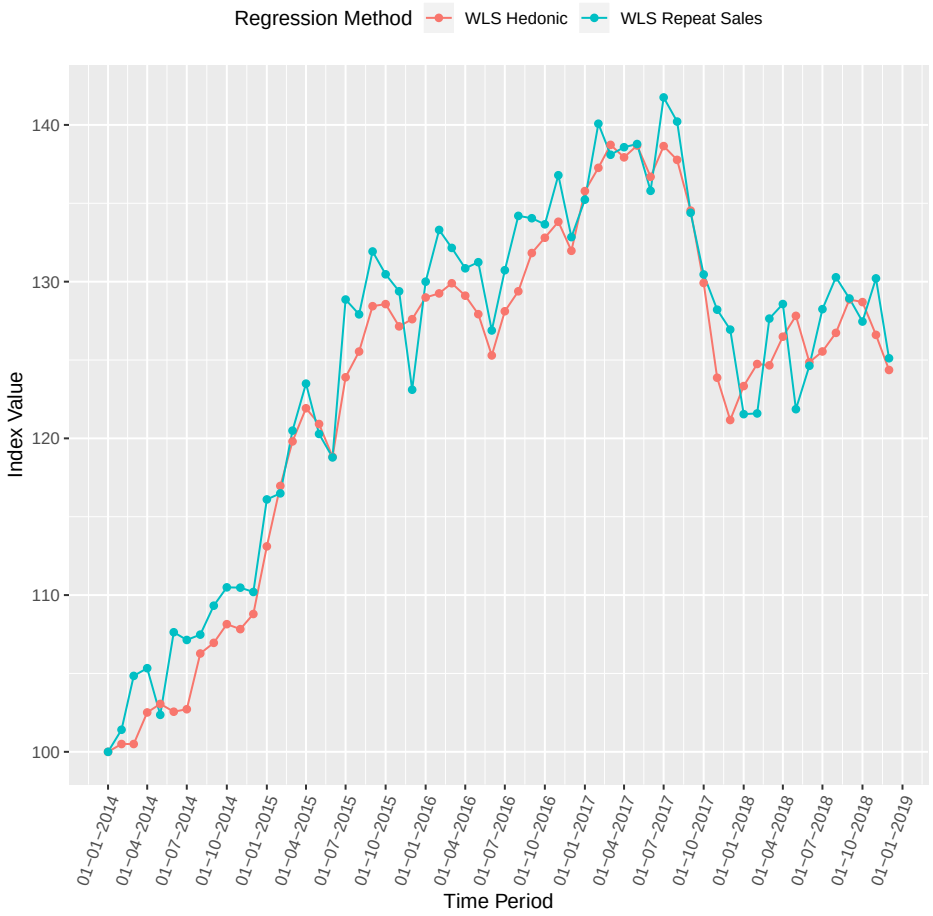

Figure 11: Weighted Least Squares; Hedonic/Repeat Sales

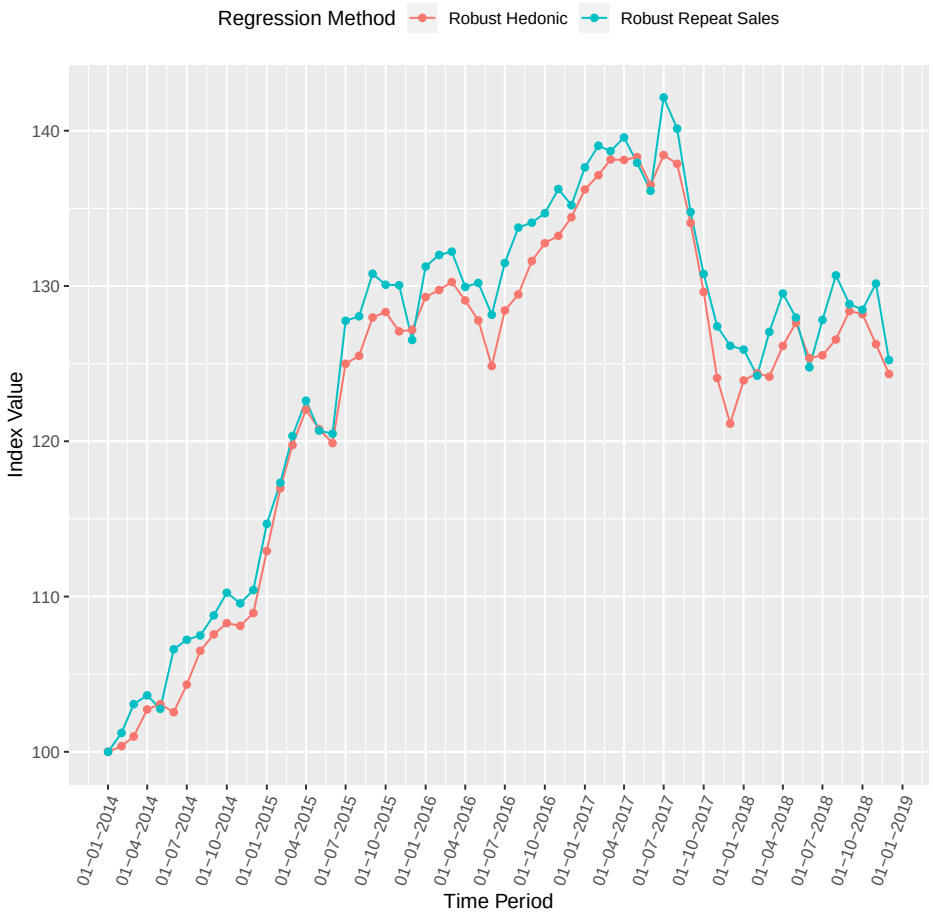

Figure 12: Robust Regression; Hedonic/Repeat Sales

7 References

1. Andersson, P., Lindensjö, K., & Tyrcha, J. (2019). Notes in Econometrics. Stockholm University: Department of Mathematics.

2. Bailey, M., Muth, R. Nourse (1963) A Regression Method for Real Estate Price Index Construction, Journal of the American Statistical Association, 58:304, 933-942

3. Bourassa, S. C., Hoesli, M., & Sun, J. (2004). A Simple Alternative House Price Index Method. SSRN Electronic Journal. doi: 10.2139/ssrn.631008

4. Case, K., & Shiller, R. (1987). Prices of Single Family Homes Since 1970: New Indexes for Four Cities. doi: 10.3386/w2393

5. Clapham, E., Englund, P., Quigley, J. M., & Redfearn, C. L. (2004). Revisiting the Past: Revision in Repeat Sales and Hedonic Indexes of House Prices. Real Estate Economics, 34(2), 275?302. doi: 10.1111/j.1540-6229.2006.00167.x

6. Clapp, J. M., & Giaccotto, C. (1999). Revisions in Repeat-Sales Price Indexes: Here Today, Gone Tomorrow? Real Estate Economics, 27(1), 79?104. doi: 10.1111/1540-6229.00767

7. Colonescu, C. (2016). Principles of Econometrics with R . Bookdown.

8. Diewert, E. (2005). Weighted Country Product Dummy Variable Regressions And Index Number Formulae. Review of Income and Wealth, 51(4), 561?570. doi: 10.1111/j.1475-4991.2005.00168.x

9. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ? Lautenbach, S. (2012). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27?46. doi: 10.1111/j.1600-0587.2012.07348.x

10. Eurostat, International Labor Organisation, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations Economic Commission for Europe and the World Bank (2013), Handbook on residential property prices indexes (RPPIs), European Union, Luxembourg.

11. Fox, J., Weisberg, S., (2011). An R companion to applied regression: Robust Regression in R. Los Angeles: SAGE.

12. Goodhart, C. and B. Hofmann (2007), Financial Conditions Indices, in House Prices and the Macroeconomy: Implications for Banking and Price Stability, Charles Goodhart (ed.), Oxford: Oxford University Press.

13. Held, L., & Bové Daniel Sabanés. (2014). Applied Statistical Inference Likelihood and Bayes. Berlin, Heidelberg: Springer Berlin Heidelberg.

14. HOXFLATSTO, Valueguard Flats Stockholm, (2020). Retrieved from http://www.nasdaqomxnordic.com/index/historiska_kurser?Instrument =SE0003077577

15. HOX Sverige. (2020). Retrieved 2020, from https://valueguard.se/indexes

16. HOXTM.Nasdaq OMX Valueguard-KTH Housing Index (2010). Retrieved from https://valueguard.se/static/media/HOX_Factsheet.e5b44ca 3.pdf

17. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning: with applications in R. New York: Springer.

18. Krause, A. (2019). A Machine Learning Approach to House Price Indexes, Seattle: Zillow Group.

19. Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models. Chicago: IRWIN.

20. Martin, J., de Adana,D. D. R., & Asuero, A. G., (2017). Fitting models to data: residual analysis, a primer, Uncertainty Quantification and Model Calibration, 133 INTECHOPEN, London

21. Pakes, A. (2003). A Reconsideration of Hedonic Price Indexes with an Application to PC?s. American Economic Review, 93(5), 1578?1596. doi: 10.1257/000282803322655455

22. Rao, C. (1973). Representations of best linear unbiased estimators in the Gauss-Markoff model with a singular dispersion matrix. Journal of Multivariate Analysis, 3(3), 276?292. doi: 10.1016/0047-259x(73)90042-0

23. Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82(1), 34?55. doi: 10.1086/260169

24. Silver, M. (2016). How to Better Measure Hedonic Residential Property Price Indexes. IMF Working Papers, 16(213), 1. doi: 10.5089/ 9781475552249.001

25. Sundberg, R. (2016). Lineära Statistiska Modeller. Stockholm: Stockholms Universitet.

26. S&P Dow Jones Indices LLC, S&P/Case-Shiller U.S. National Home Price Index [CSUSHPISA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CSUSHPISA, March 29, 2020.