Analysis of the Regional Purchasing Power in the Swedish Housing Market

A Multiple Regression Modeling Approach

THERESE LIN

STEFAN MARAS
Analysis of the Regional Purchasing Power in the Swedish Housing Market
A Multiple Regression Modeling Approach

THERESE LIN
STEFAN MARAS
Abstract

This report analyzes which macroeconomic factors have the greatest effect on a consumer’s purchasing power on the Swedish housing market. Multiple linear regression analysis is used to construct separate models for each of the eight Swedish regions: West Sweden, South Sweden, Småland and the Islands, Stockholm, East Middle Sweden, Middle Northern Sweden, North Middle Sweden and Upper Northern Sweden. The chosen factors, which are also used as regressors in the regression model, are disposable income, increase in households, employment rate, the Riksbank’s repo rate, square meter price for tenant owned flats and new construction of residential homes. Quarterly data is collected and processed for every macroeconomic factor from the second quarter of 2005 to the fourth quarter of 2015. The affordability is defined as the consumer closing price for tenant owned flats and it is used as the response variable for the regression models. The models are first fitted using all the regressors. To improve the models and potentially eliminate some of the macroeconomic factors a number of different model adjustments, such as Cook’s distance, VIF evaluation, stepwise forward and backward elimination are used. The results show that the purchasing power among the regions is mainly described by disposable income, employment rate and increase in households. Disposable income is the macroeconomic factor that has the greatest effect on the regional purchasing power.
Sammanfattning

Acknowledgements

We would like to extend a special thanks to our supervisor at the department of mathematics at KTH, Tatjana Pavlenko, for her support in discussing and evaluating the writing process of the report. Additionally, we would like to thank Susanne Spector and Viktoria Oleson, and their associates at the department of Strategy and Research at Nordea for their help in choosing relevant variables and problem formulations concerning the regional analysis.
# Background

## Purpose and problem formulation

## Limitations

## Theoretical background

### Economic terms

- **Purchasing power**
- **Consumer closing price**
- **Inflation**
- **Riksbank’s repo rate**
- **Income**
- **New construction of residential buildings**
- **Employment rate**
- **Square meter price**

### Regression analysis

- **The Multiple Regression Model**
- **Ordinary Least-Squares Estimation**
- **Key assumptions of the OLS estimator**
- **Multicollinearity**
- **Variance Inflation Factor**
- **R²**
- **Adjusted R²**
- **F-statistic and P-value**
- **Mallows’ Cp**
- **Bayesian Information Criterion (BIC)**
- **Cook's Distance**
- **Best subsets regression**
- **Stepwise forward selection**
- **Stepwise backwards selection**
- **k-fold cross-validation**

## Data analysis

### Data preprocessing

### Regression analysis

### Methodology of the results

## Results

### West Sweden

### South Sweden

### Småland and the Islands

### Stockholm

### East Middle Sweden
North Middle Sweden 32
Middle Northern Sweden 34
Upper Northern Sweden 37

Discussion 41
West Sweden 41
South Sweden 41
Småland and the Islands 41
Stockholm 41
East Middle Sweden 42
North Middle Sweden 42
Middle Northern Sweden 42
Upper Northern Sweden 42
Macroeconomic Factors 44
Purchasing power 45
Data 45
Future analysis 45

Conclusion 47

References 48
Background

The Swedish housing market has been a growing market for a long time, influenced by factors such as regional regulations, income distribution, new construction of residential buildings and demographic changes. Today it is facing an increased demand and a varying purchasing power among its population, however, the housing market has not always been the way that it is today. It is thus interesting to evaluate the late history of the Swedish housing market by looking at how key macroeconomic factors affect it and the consumers’ purchasing power in it. In the context of this paper, the purchasing power among consumers is referred to as a household’s ability to buy/exchange housing property on the Swedish housing market. In other words, this concept describes a financial possibility, based on one’s economic situation and the surrounding socio-economic structure, to finance and own property.

The first macroeconomic factor to consider is the average disposable consumer income. It has been steadily increasing from the 1990’s despite the economic pressure from the financial crisis’ of that time. The increase was impaired after the financial crisis of the early 1990’s due to large government austerity measures. These developments are seen to be paralleled with the housing price index of the time which took a fall during the financial crisis between 1990-1993, only to then increase linearly from 1994 onward. A correlation can thus be seen between consumers’ disposable income and the development of the country’s housing prices. As housing prices increase so does the transaction demand on money and income as consumers want to maintain an equivalent purchasing power to the increased prices.

The housing prices can also be explained by looking at the square-meter-price of different residences in the selected regions. The increasing price per square meter trend can be seen to follow the same pattern as the income and housing prices. There is for example a clear drop in price around 2007-2008 which can be explained by the financial crisis of the time. The square-meter-price thus acts as a good historic estimator of housing prices since the total residence value will be based on the possible living area.

The previously stated premise is also apparent when considering the official bank rate in Sweden as it affects the leverage of property owners (and therefore the housing prices implicitly) as well as the purchasing power of the consumer. Contrary to the previous factors, historical values show a decreasing official bank rate in Sweden which might explain the other historic pricing differences by correlation.

Historically housing prices have been affected by the supply of buildings and accommodations. This implies that newly built homes, which increases supply, will influence the prices on the housing market. Sweden has seen a fluctuating new-production of buildings. During the time period from 1990 to 2014 Sweden saw a large decrease in newly built buildings in the early 90’s, but afterwards, the
production stabilized and increased until the financial crisis of 2008 where the housing market faced another decline in new-production. Since then, the quantity of newly built condominiums and tenancies have steadily increased.\(^5\) A parallel to this could be seen in the increase of households, i.e. population growth divided by two.

The increase-in-household factor, it can be seen as an implicit reflection of the demand on the housing market, affecting its prices up or down depending on the state of the market. Historically, since Sweden’s population has been steadily growing since the 1950’s, so has the increase in households\(^6\). The effect of this increase of households would intuitively imply an increase in prices since more people demand housing accommodations.

The last factor to be introduced is the employment rate which can be used in different ways to describe the economic climate in Sweden. The historical development of this factor has been volatile, but using a method such as the least-squares estimate would yield a positive slope indicating a historical net increase.\(^7\) This factor is interesting since people usually move to where there are more jobs, i.e. areas which will have a higher employment rate. This could therefore be an indicator of how housing prices are affected by a larger (in proportion to an unemployed) workforce.


Purpose and problem formulation

The aim of this paper is to analyze the purchasing power on the Swedish housing market. The housing market is important for the development of a society and can have an impairing as well as a driving force on the economy. Concerning Sweden’s geography, it is common to look at the country by dividing it into 8 characteristic regions; West Sweden, South Sweden, Småland and the Islands (Stockholm, East Middle Sweden, North Middle Sweden, Middle Northern Sweden and Upper Northern Sweden. These regions are driven by common as well as differing economic conditions which together comprise the Swedish housing market. By analyzing the market on a regional basis, the Swedish housing market can be broken down and thoroughly reviewed. By going through each region and evaluating its historical development, new future forecasts on the market can be developed using multiple linear regression analysis.

The following problem specification has been posed to examine the above stated premise:

- Which macroeconomic factors have the greatest effect on the consumer’s purchasing power on the Swedish housing market, examined on a regional level?

This report is written in association with Nordea markets. The problem formulation was given by the department of Strategy and Research at Nordea.
Limitations

The greatest limitation imposed on the project’s premise is that of the geographical division of Sweden into eight regions. No geographical areas outside the Swedish borders will be used when conducting the analysis. The report will therefore exclude any foreign purchases of Swedish real estate and/or housing.

The direct and indirect macroeconomic factors contributing to the markets historical and future development are in surplus. In the context of a regression analysis, a large amount of regressor variables, e.g. economic factors, have the possibility of causing problems to the model, such as contributing to the rise of multicollinearity. To construct the multiple linear regression analysis, the six variables mentioned in the background have been chosen. These are the: Disposable income, Square-meter price, the Riksbank’s repo rate, New construction of residential homes, Increase in household, and the employment rate. By choosing these factors as variables, a concise selection of information can be used to evaluate the subject matter in a condensed form.

The time frame of the data will be restricted to the time period from 2005 to 2015. Data from prior 2005 was not available for all the macroeconomic factors. The amortization requirements that were introduced in 2015 have significant impact on a household’s ability to take mortgage loans and purchase homes. The time period after 2015 were therefore excluded.

Furthermore, the average disposable income gathered will be limited to the private sector. This is because this sector houses the largest part of the workforce in Sweden and is often characterized by a larger net income.

---


Theoretical background

For the sake of clarity, it is important to establish key terms as well as their meaning and purpose in the context of this case study.

Economic terms

Purchasing power

The purchasing power is one the most important and fundamental terms in this paper. Purchasing power is defined as a household’s ability to buy/exchange on the Swedish housing market. Affordability is in practice the equivalent term to purchasing power, it describes what a household can afford on the housing market, and specifically, the Swedish housing market. When economic evaluations of households’ affordability are made, their economic positions are based on indices and percentiles which try to quantify what households can do with respect to their demographic backgrounds. In this context, the two indices which are of interest to the subject matter of this report are the Housing Affordability Index (HAI) and the Housing Opportunity Index (HOI).

The HAI rates housing affordability where a value of 100 means that a family with the median income has exactly enough income to qualify for a mortgage on a median-priced home. The HOI on the other hand measures the share of homes sold in an area that would have been affordable to a family earning the local median income, based on standard mortgage underwriting criteria.

Consumer closing price

In this study the consumer closing price will define the purchasing power for households in the different Swedish regions. The closing price tell how much the households spend on their homes and thereby show the households affordability.

Figure 1. The figure below shows the consumer closing price of homes in the Swedish regions from 2005-2015.

![Figure 1 - Consumer Closing Price](image_url)
Inflation

Historically, the prices on housing property in Sweden have been rising. When the price of an individual good increases in value it is called a relative price increase. There could be many reasons for this, for example, the demand on housing increases, which makes the prices higher. Another is inflation, which is the rate at which the general level of prices for goods and services increase. The price increase is lasting, and the value of money is declining. Inflation is therefore also a responsible factor in the increasing prices of housing property and may therefore affect the purchasing power among consumers.12

There are several reasons behind inflation. When the central bank supplies too much money, prices rise, and the value of money is undermined. Inflation can also be caused when the demand on the market is bigger than the supply. The cost of producing goods and services increases leading to higher prices and inflation.13

High and fluctuating inflation can be problematic as the price development is uncertain. Households and companies encounter difficulty when planning to take financial decisions. The willingness to invest declines and the activity in the economy slows down. If inflation is too low the risk of deflation (fall in the general price level) rises. However, a low and stable inflation can create good conditions for a favorable economic growth. Many central banks around the world focus on keeping the inflation low and stable.14

One way to measure inflation is to use a consumer price index, CPI. However, the official bank rate affects household’s mortgage rates more prominently. This affects CPI as the household’s living costs also change. Therefore, inflation is sometimes measured using consumer price index with a fixed interest rate, CPIF, to eliminate this effect. The statistical bureau, SCB, is responsible for official and governmental statistics in Sweden. SCB collects Swedish price data for many goods and services every month and derives an average price level for the current cost of living compared to the previous month to get the CPI. CPIF is calculated using the same data as CPI, however the changes to mortgage rates are excluded.15

The Swedish Riksbank’s inflation target was introduced in 1993 and began to apply from 1995. Today the target is to hold inflation around 2 percent a year in terms of the CPIF. Since September 2017, Riksbank uses a variation band of 1 - 3 percent for the outcomes for inflation. The target functions as a benchmark to guide the future economic development and helps households and companies to make economic decisions.16

---

13 Ibid
Riksbank’s repo rate

The official bank rate has been the Riksbank’s official policy rate since 1994. It is the rate of interest at which commercial banks can borrow or deposit funds at the Riksbank for a period of seven days. When the official bank rate increases the commercial banks must pay higher interest to the central bank. To compensate, the commercial banks raise their interest rate on the loans they offer to customers. As the lending rate increases, the money becomes more expensive and the borrowing of new money decreases. In cases when inflation is too high it is beneficial to increase the official bank rate as the money supply in the economy is reduced. However, when the official bank rate decreases the money supply in the economy increases. These spikes increase in demand of goods, household investments like housing, and a positive GDP growth.

The Riksbank introduced a negative official bank rate in 2015. Today the official bank rate is set to -0.50 %. As mentioned in the previous section the inflation target is 2 percent. Swedish inflation has been close to the target, however failed to meet the exact goal. The record-low rate is set to stabilize close to the target going forward.

Figure 2. The table below shows the Riksbank’s repo rate from 2000 - 2016

![Figure 2 - Riksbank’s Repo Rate](image)

**Income**

High nominal wages do not necessarily mean higher purchasing power since they do not consider the changes in price level. The real wage is the nominal wage adjusted for inflation. A high wage growth can stimulate the economy and contribute to higher inflation. For example, in 1970 - 1995 the nominal wage growth was high, in average the wage increased with 8.1 percent each year. However, the high wage growth boosted the inflation to an average of 7.6 percent each year. This resulted in an actual growth in real wages of only 0.5 percent per year. During the period 1995 - 2016 the average nominal wage growth was 1.8 percent per year.

---

18 Fodérus, Leo; teacher at the department of industrial engineering and management, Kungliga Tekniska Högskolan. Stockholm, lecture 5, 2018-01-30
wage growth was 3.4 percent. However, the low inflation gave an average real wage growth of 2.4 percent each year.\textsuperscript{21}

The last couple of years have had a lower growth in nominal wage of 2.5 percent. According to Nordea’s forecast the nominal wage growth will be a bit over 3 percent in 2019.\textsuperscript{22}

Figure 3. The table below show the income in the Swedish regions from 2005 - 2016.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{ Disposable_income.png}
\caption{Disposable Income}
\end{figure}

New construction of residential buildings

The Swedish banking rescue 1990-1994 and the tax reform in 1993 resulted in a decrease of new construction in Sweden. During the last couple of years there has been an increase in new construction, mainly condominiums and houses. A study from 2017 made by SBAB showed that the market offered more newly constructed homes than existing homes. In June 2017 there were 14 781 new constructions and 13 271 existing homes for sale.\textsuperscript{23}

\begin{itemize}
\item \textsuperscript{21} Carlgren, Fredrik. 2017. Real lönutveckling i Sverige. \textit{Ekonomifakta}. \url{https://www.ukonomifakta.se/Fakta/Arbetsmarknad/Loner/Loneutveckling-och-inflation/} (Accessed 2018-04-06)
\end{itemize}
Figure 4. The table below show the new construction of homes in the Swedish regions from 2005 - 2017.

![New construction of homes](image)

**Figure 4 - New Construction of Homes**

**Employment rate**

The employment rate is an important macroeconomic indicator that show the state of the economy. The aim is to have a high employment rate. The employment rate is defined as the percentage of the labor force that is employed. ²⁴

For the defined time period the employment rate reached its lowest point in 2009-2010 and its highest in 2008 and 2017.

Figure 5. The table below show the employment rate in the Swedish regions from 2005 - 2017.

![Employment rate](image)

**Figure 5 - Employment Rate**

Square meter price

The square meter price is the average price for one square meter housing. From the data we can see that the square meter price has steadily been growing since 2005 to 2017. This data resembles the data for the consumer closing price.

Figure 6. The table below show the square meter price of homes in the Swedish regions from 2005 - 2017.

Regression analysis

The Multiple Regression Model

Regression analysis is used to understand how independent variables are related to a dependent variable and find the relationships between the regressors and the response variable. When there are more than one regressor variable it is called a multiple regression model. The structure of the multiple regression model is as follows, in matrix notations:

\[ Y = X\beta + \epsilon \]

where \( Y \) is an \( n \times 1 \) vector containing the \( n \) observations, i.e. the response variable, \( \beta \) a \( p \times 1 \) vector containing the regressor coefficients, \( X \) an \( n \times p \) matrix containing all the \( n \) values from all the \( p \) regressor variables, and finally \( \epsilon \) an \( n \times 1 \) vector of random errors from the data, these error terms are assumed to be normally distributed and uncorrelated.\(^{25}\) Another key assumption that is made is that the residuals are derived from a Gaussian distribution, this is assumed to ensure that the future F-tests are possible to perform and the F-statistics accurate.

Ordinary Least-Squares Estimation

Since the regressor coefficients are unknown, they have to be estimated. There are several methods available but the one used here will be the most commonly applied on the basic multiple linear regression model, that being ordinary least square (OLS) estimation. The principal of the OLS estimation is to estimate the regression coefficients so that the sum of squares of the difference of the observations $Y$ and the lines $X\beta$, i.e. the residuals, is a minimum. It is thus desired to find the vector of OLS estimator, which minimizes:

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = (Y - X\beta)^T (Y - X\beta)$$

$$= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta$$

The OLS estimator must satisfy:

$$\frac{\partial S(\beta)}{\partial \beta} \bigg|_{\beta = \hat{\beta}} = -2X^TY + 2X^TX\hat{\beta} = 0$$

which then simplifies to:

$$X^TX\hat{\beta} = X^TY$$

which in turn finally yields the OLS estimation:

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

The regression models can be used to predict future data as well as fill in missing data points for the chosen response variable.

Key assumptions of the OLS estimator

OLS estimation makes several key assumptions to perform model adjustments and fitting. The assumptions that are made and which stand relevant for the regression analysis of this paper are:

1. There must be a linear relationship between the response $Y$ and the regressor variables, at least approximately.
2. The error terms are assumed to be normally distributed with an expected value of 0 and a constant variance of $\sigma^2$.
3. The errors are assumed to be uncorrelated. Violation of this criteria would imply that the OLS estimators are not suitable.
4. The correlation matrix $(X^TX)^{-1}$ exists, i.e. the columns of $X^TX$ are linearly independent, there is thus no severe multicollinearity present.
5. The regressors are fixed variables with no error in measurement.

Multicollinearity

Multicollinearity implies that two or more of the regressors are moderately/highly correlated, i.e. there exists near-linear dependence among the regressors. Multicollinearity is a serious problem that may limit the analysis of the dataset used and affect the performance of the constructed model.  

---

There are four different primary sources for multicollinearity: the method used when collecting data, the constraints, the model specifications and an over defined model.\textsuperscript{27}

The employed method when collecting data can be problematic when only a subspace of the region of the regressor defined is sampled. The sampling technique is not inherent in the model and there is usually nothing in the physical structure of the problem to prevent this.\textsuperscript{28}

Some regressors will have a natural correlation. For example, high income households generally have larger homes than low income households. This kind of natural constraints on the model or in the model being sampled will exist regardless of the sampling method employed. This cause of multicollinearity is usually found in problems involving processes where the regressor are the components of a product which add to a constant.\textsuperscript{29}

The choice of model can contribute to multicollinearity. For example, if the range of x is small, adding a $x^2$ term can cause stronger correlation between the variables. Retaining all regressors when they are linearly dependent may induce a serious problem for the model. Usually some subset of the regressor variables is preferable to minimize multicollinearity.\textsuperscript{30}

When there are more regressor variables than observations, an over defined model is constructed. The information used in the regression is collected for a too large number of regressor variables when there are only small sample units for each variable. Variable elimination is the most effective way of solving multicollinearity caused by this.\textsuperscript{31}

The data of the regressors form the columns of the $X$ matrix, which means that a singular $X^T X$ matrix implies exact linear dependence.\textsuperscript{32} The regression coefficients are inflated if there exists multicollinearity. If the off-diagonal elements are not equal to zero, the regressors are not orthogonal and multicollinearity is evident.\textsuperscript{33}

**Variance Inflation Factor**

The diagonal elements of the $C = (X^T X)^{-1}$ matrix, i.e. the inverted correlation matrix, are called the variance inflation factors (VIF.s) and can be used to detect multicollinearity.

The VIF for the $j$:th regression coefficient can be written as:

$$VIF_j = \frac{1}{1-R_j^2}, \quad VIF_j = C_{jj} = (1 - R_j^2)^{-1}.$$
where $R_j^2$ is the coefficient of determination that is obtained when $x_j$ of the $X$-matrix is regressed with respect to the remaining $p-1$ regressor variables.

The VIF, as a measure, quantifies for each regressor in the model, the effect on a regressor coefficient’s variance because of collinearity among the regressors. A common rule is that if any VIF exceeds a value 5 and/or 10 then this in an indication that the problem of multicollinearity exists and that the regressor coefficients are poorly estimated.\textsuperscript{34}

**$R^2$**

This term is called the coefficient of determination and defined as:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

where $SS_R = \text{Regression Sum of Squares} = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$

and $SS_{Res} = \text{Residual Sum of Squares} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$,

where $y_i$ are the observed values, $\hat{y}_i$ the fitted values $= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ where $x_{ij}$ is the $j$th element of the $i$th row of the $X$ matrix), and $\bar{y}$ the mean value of the observed values $= \frac{1}{n} \sum_{i=1}^{n} y_i$.

The $R^2$-value is a measure of the variability in y that is explained by the regression model, e.g. the predictive ability of the fitted model. Values can range from 0 to 1, with 1 meaning that theoretically all variability in y can be explained by the regression model. The measure should however be used with caution since it possible to make the $R^2$ statistic large by adding enough terms to the model. For this reason, the adjusted $R^2$ value is more commonly used when performing multiple linear regression, something that is also done in this report when comparing different models from the following results.\textsuperscript{35}

**Adjusted $R^2$**

The adjusted $R^2$ value follows the same premise as the original $R^2$-statistic but with weighted sum of square variables resulting in the following definition:

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T(n-1)},$$

where $n$ = number of observations and $p$ = number of regressor variables. From this definition, it follows that the $R^2_{Adj}$-statistic will only increase if the addition of a regressor variable reduces the residual mean square, i.e. the $SS_{Res}/(n-p)$-term is reduced. The $R^2_{Adj}$-statistic thus penalizes the addition of too many variables if they do not contribute to the regression model which makes it a better measure of the regression model’s predictive ability in the multiple linear case compared to the original $R^2$-statistic.\textsuperscript{36}


F-statistic and P-value

The F-test is an analysis-of-variance test to evaluate if the variance between the regressors in the regression model is due to random chance, or not. Rejecting the null hypothesis is equivalent to at least one of the regressors having a significant effect on the response variable. The test is set up as follows:

\[ H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0, \quad H_1: \beta_j \neq 0 \]

\[ F = \frac{SS_R/df_R}{SS_{Res}/df_{Res}} \]

where \( SS_R \) and \( SS_{Res} \) are the same quantities as defined in the R²-parapgraph, while \( df_R \) = degrees of freedom in the regression model = \( p \), and \( df_{Res} \) = Degrees of freedom among the residuals = \( n - p - 1 \), where \( n \) = number of observations.\(^{37}\) To test the null hypothesis, the test statistic \( F \) is calculated and \( H_0 \) rejected if \( F > F_{\alpha,p,n-p-1} \) where \( \alpha \) is the quantified quantile based on the desired significance\(^{38}\), e.g. if a significance of 95% is desired then \( \alpha = 0.05 \).

The p-value is the probability, when the null hypothesis holds true, that a statistical summary, in this case the F-statistic, would be of the same significance, or of greater significance, than the actually observed results.\(^{39}\) This implies that a smaller p-value, one smaller than the alpha level, is an indicator of the rejection of the null hypothesis.

Note: The F-test and -statistics are calculated in R-studios and evaluated externally.

Mallows’ Cp

Mallows’ Cp is a statistic used as criteria for deciding the optimal number of variables in a regression model. The statistic is defined as:

\[ C_p = \frac{SS_{Res}(p)}{\hat{\sigma}^2} + n - 2p, \]

where \( \hat{\sigma}^2 \) is a good estimate of the variance of the fitted values, \( n \) = the number of observations in the data, and \( p \) = the number of regressors in the fitted model.\(^{40}\)

In general, small values of \( C_p \) are desirable and its values are usually plotted as a function of \( p \) to visualize the optimal number of regressors to use in the model.\(^{41}\)

Bayesian Information Criterion (BIC)

The BIC is a Bayesian extension of the Akaike Information Criterion (AIC) which in essence is a penalized log-likelihood measure. The principle of the AIC is grounded in maximizing the expected information of the regression model. In the case of the BIC, it is defined as:

\[ BIC = 2\ln(L) + p \star \ln(n) \]


\(^{38}\) Ibid


\(^{40}\) Montgomery C., Douglas et al.. Introduction to Linear Regression Analysis. 5. edition. John Wiley & Sons, Inc.. 2012. Page 335

\(^{41}\) Ibid
where \( L \) = the likelihood function of a specific model, \( p \) = the number of regressors in the model, and \( n \) = the number of observations.\(^{42}\) The difference in the criteria is that the BIC penalizes the addition of regressors as the sample size increases more severely than the AIC. When performing analyses using this criterion, the model with the lowest BIC-value is chosen as the optimal final model.\(^{43}\)

**Cook’s Distance**

When evaluating data sets it is desirable to consider both the location of a point in the x space and the response variable in measuring influence. It is thus calculated using a measure of the squared distance between the least-squares estimate (based on all \( n \) observations) and the estimate obtained by deleting the \( i \)th point, say \( \hat{\beta}_{(i)} \). This distance measure can thus be expressed in a general form as:

\[
D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{p MS_{Res}} = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}
\]

where \( p \) = the number of regressor variables in the model, \( MS_{Res} = \) the residual mean-square error = \( \frac{SS_{Res}}{n-p-1} \), \( r_i \) = the \( i \)th studentized residual = \( \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}} \), \( h_{ii} \) = the diagonal elements of the hat matrix \( H \) defined as \( H = X(X^T X)^T \). \( h_{ii} \) can thus be interpreted as the distance from the remaining data points and expresses the effect an observation \( y_i \) has on a fitted value \( \hat{y}_i \). Finally, the ratio \( h_{ii}/(1-h_{ii}) \) can be shown to equate the distance of the vector \( x_i \) from the centroid of the remaining dataset. \( D_i \) thus combines the residual magnitude for the \( i \)th observation and the location of that point in x-space to assess influence. This means that a large \( D_i \) usually implies the existence of an influential point.\(^{44}\)

**Best subsets regression**

This algorithm identifies all of the possible regression models from all of the possible combinations of the candidate regressor variables. If there are \( k \) candidate regressors, then there are \( 2^k \) possible regression models that can be considered.\(^{45}\)

By going through the possible models identified in the first step, the one, two, and so on predictor models that do the best at meeting some well defined criteria are determined and extracted. These criteria can be different depending on what the regression model is set to analyze but the most common criteria are the adjusted \( R^2 \) (penalizes the addition of more regressors to a model), mean-square error (MSE) (quantifies how far away our predicted responses are from our observed responses) and Mallows’ \( C_p \)-statistic (estimate of the size of the bias that is introduced into the predicted responses by having an underspecified model).\(^{46}\)

---


\(^{43}\) Ibid


\(^{46}\) Ibid
Stepwise forward selection
This algorithm is described in An Introduction to Statistical Learning by beginning “ [...] with the null model - a model than contains an intercept but no predictors. We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS [Residual Sum of Squares]. We then add to that model the variable that results in the lowest RSS for the new two-variable model. This approach is continued until some stopping rule is satisfied”.47

Stepwise backwards selection
This algorithm is described in An Introduction to Statistical Learning by beginning “ [...] with a variable in the model, and remove the variable with the largest p-value - that is, the variable that is the least statistically significant. The new (p - 1)-variable model is fit, and the variable with the largest p-value is removed. This procedure continues until a stopping rule is reached. For instance, we may stop when all remaining variables have a p-value below some threshold”.48

k-fold cross-validation
This algorithm is described in An Introduction to Statistical Learning by “[...] randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k - 1 folds. The mean squared error, MSE_i, is then computed on the observations in the held-out fold. This procedure is repeated k times; each time, a different group of observations is treated as a validation set. This process results in k estimates of the test error, MSE_1, MSE_2, ..., MSE_k. The k-fold cross-validation estimate is computed by averaging these values, ”

\[
CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i
\]

The model with the smallest error is then chosen.49

---

47 James, Gareth et al.. An Introduction to Statistical Learning with Applications in R. Springer-Verlag New York, Inc.. 2013. Page 78
48 James, Gareth et al.. An Introduction to Statistical Learning with Applications in R. Springer-Verlag New York, Inc.. 2013. Page 79
49 James, Gareth et al.. An Introduction to Statistical Learning with Applications in R. Springer-Verlag New York, Inc.. 2013. Page 181
Data analysis

The basis for this report lies in a literature study where analysis of the Swedish housing market is made to work as a platform from which new results and conclusions can be drawn through a quantitative study using multiple linear regression. The regression modelling is performed in R studio, using the programming language of R. R is used mainly for statistical computing and graphics, which will also be used for this report.

It was previously mentioned that the most effective way of evaluating the Swedish housing market is by dividing it into eight characteristic regions. When applying a linear model on this division there are two approaches which can be used. Produce one model for the whole country where each region acts as an individual factor to the complete model, or develop separate models for each region, i.e. eight models based on the region-specific data, gathered from SCB and Macrobond. The choice for this report is to develop eight separate models since this will contribute to a more accurate depiction of the purchasing power across the country.

Data preprocessing

The data which is to be used has been compiled in separate excel documents, one for each region. For each factor: consumer closing price, disposable income, official bank rate, newly built homes, square meter price and change in household, the observations are in the form of quarters as follows:

Q1: 1 January - 31 March
Q2: 1 April - 30 June
Q3: 1 July - 30 September
Q4: 1 October - 31 December.

A table of the Riksbank’s monthly official bank rate could be found on Ekonomifakta. The official bank rate is set by the central bank of Sweden and is therefore the same in all regions.50

The data found in Swedish Statistics, SCB, database was for the Swedish counties, not the regions we defined earlier in this report. Therefore, the data needed to be reconstructed by adding the data from different counties together as shown below:

Middle Northern Sweden: Västernorrlands län + Jämtlands län
North Middle Sweden: Värmlands län + Dalarnas län + Gävleborgs län
Småland and the Islands: Jönköpings län + Kronobergs län + Kalmar län + Gotland län
Stockholm: Stockholms län
South Sweden: Blekinge län + Skåne län
West Sweden: Hallands län + Västra Götalands län
East Middle Sweden: Uppsala län + Södermanlands län + Östergötlands län + Örebro län + Västmanlands län
Upper Northern Sweden: Västerbottens län + Norrbottens län

On average, each household consists of 2 people.51 To get the change in household for each region the population growth in number of citizens in each region is divided by two. Since we only had yearly

data for 2005 - 2015, the yearly change in household is divided by four to get the change for each quarter. Quarterly data for 2016 and 2017 could be found in Swedish Statistics database.\textsuperscript{52}

To get the employment rate for each region the sum of employed in the age group of 15 - 74 years old is divided by the total population in the same age group.

\[
\text{Employment rate} = \frac{\sum \text{Employed 15 – 74 y/o}}{\sum \text{Population 15 – 74 y/o}}
\]

To construct the data for the income linear interpolation is used. The average change for a year is calculated by dividing the yearly change by four and then added to the

\[
c = \text{income in the start of the year} \\
ap = \text{average change per quarter.} \\
T = \text{quarter (1,2,3, or 4).} \\
Z = \text{income at quarter.}
\]

\[
Z = ap + c
\]

Data for ending price, new construction of homes, square meter price exists for each quarter of the defined period in SCB’s database.

Regression analysis

A very important part of modelling data using regression analysis is to determine the response and regressor variables. The purpose of the report was to identify the macroeconomic factors which affects the purchasing power of the individual on the Swedish housing market the most. The background thus introduced the regressor variables: Average disposable income, Square-meter price, Official bank rate in Sweden, New built estate, Increase in household, and the Employment-to-population ratio. These are relevant since they individually affect the housing prices on the Swedish housing market while staying relatively independent of each other.

Concerning the response variable, the obvious choice would be the housing prices specific to each region. However, after discussion with Nordea analysts the more accurate choice of a response variable would be the consumer closing price, i.e. the final price of the house/apartment that the consumer pays. This value can be assumed to be normally distributed since prices vary up and down depending on the geographical location of the house/apartment as well as the demographic of the population demanding these accommodations.

The full model thus becomes: Consumer Closing Price ~ Average Disposable Income + Square-Meter Price + Official Bank Rate of Sweden + New Built Estate + Increase in Household + Employment-to-population ratio. The data used to comprise this model will in turn be gathered from the previous 11 years starting from 2005 to 2015 and be read quarterly, which in turn will yield 43 (not 44, since the

\textsuperscript{51} SCB and Macrobond

first quarter is excluded) data points per region. This choice of time span and division in quarters is reasonable since an accurate regression model demands a fair amount of observations to accurately produce reliable models while maintaining a reasonable data set. If a division would have been made monthly than this would imply 128 observations per region which, under consideration, seemed superfluous and was there decided against.

After the model has been derived and fitted several different model adjustments will be made to evaluate and perfect the model. These include using datapoint selection methods such as Cook’s and normally distributed plots to find and assess potential outliers and leverage points. Furthermore, variable selection algorithms will also be implemented to dissect which regressor variables are the most influential and possibly remove the ones that are excessive. Among these variable selection algorithms are the Backwards and Forwards BIC elimination as well as Best Subsets Regression.

Methodology of the results

The results follow a chronological structure where, first, a summary of the full model is evaluated to check for greatness of fit as well as the significance of the individual regressors and the model as a whole. Second, a correlation matrix is studied to evaluate any possibility of multicollinearity arising from the correlation between the regressor variables. To further analyze the possibility of multicollinearity, each variable’s VIF is calculated and analyzed using a value of 5 as the cut-off. If a regressor variable is removed a new, reduced, model is fitted and used when conducting the following tests.

The following tests include residual analysis and outlier evaluation. To perform these tests the residuals are plotted after a theoretical normal distribution using the normal QQ-plot in R-studio, the residuals versus the fitted values are plotted, the Cook’s distance and the residual density plots are also plotted. While studying the normal-QQ plot and the residuals density plot it is relevant to check if the data points are close to the theoretical normally distributed line or if there is exists any skews or tails indicating the need for any transformations. In the case of a tail or skew present in the plots, the data is first transformed and then the model refitted before plotting the above-mentioned plots anew. Proceeding with the two remaining plots the residuals versus fitted plot is used to identify the possibility of a transformation by looking at the spread of the points in x- and y-space. The Cook’s distance plot is used to identify any possible outliers. As a rule of thumb, a distance 3 times greater than the mean was used as an indicator of a possible outlier. In the case of any outliers, these points are investigated by looking at the original data and identifying any outlying values from the regressors which might have affected the points reliability, e.g. one of the regressor values for the point might have been incorrectly calculated or measured. If the point is found to be incorrectly measured or is affecting some of the assumptions but not the result it will be removed from the data set.

Finally, variable selection algorithms are used on the reduced model to evaluate if there is another smaller model that might have a better predictive ability than the larger reduced model. To test this, stepwise backwards selection, stepwise forward selection, and best subset regression is used and the results between these models evaluated. The criteria used when evaluating the results from the variable selection algorithms are the residual sum of squares, the $R^2$-value, the adjusted $R^2$-value, Mallows’ Cp, and finally the BIC-value. The criteria are weighted against each other and the number of variables in the reduced model is chosen after the suggestions of each criterion. The number of regressor variables suggested is then compared with results from a k-fold cross-validation on the data.
which employs $k = 5$ folds since the dataset is small with its 43 points. If the 5-fold cross-validation coincides with the variable selection algorithms than this suggested model is compared to the previous model and a final model is decided upon based on the significance from the F-statistic, p-value, $R^2$ values, and the significance of the regressor variables. In this case, a 95% significance is considered satisfactory and in the case of an improvement being found in the smaller reduced model, this is valued secondly to the significance level. However, if the 5-fold cross-validation differs from the variable selection algorithms then both suggestions are considered and three models (as opposed to two models following the first outcome) are compared and evaluated based on the same criteria just mentioned. A final model is then presented after further evaluation done in the first section of the discussion.
Results

Since the regression analysis focuses on modelling eight regions separately, the results of the analysis will be represented region by region, with some overlapping similarities. In what follows, when displaying the final model, R syntax will be used: \( y \sim x_1 + x_2 + \ldots + x_k \), where \( x_i \) represent the variable chosen.

West Sweden

For the full model using all the variables the F-statistic was large and the p-value below 0.001. The adjusted-R\(^2\) was above 0.8. The significance of the regressor variables were mixed with more than half having a significance lower than 95%. The correlation matrix indicated that a strong relationship existed between the income, household and square-meter price variables. The VIF-values is presented in the table below:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
<th>sqp</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>9.010049</td>
<td>2.947405</td>
<td>1.558101</td>
<td>8.786319</td>
<td>1.878899</td>
<td>18.995033</td>
</tr>
</tbody>
</table>

Table 1.1 - VIF of the Full Model (FM)

The VIF table above show three variables, same as the ones observed in the correlation matrix, with a VIF larger than 5. The Square-meter price variable had the largest VIF and was removed.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>3.696803</td>
<td>2.801520</td>
<td>1.479598</td>
<td>5.336878</td>
<td>1.798027</td>
</tr>
</tbody>
</table>

Table 1.2 - VIF of the First Reduced Model (FRM)

The household variable still had a VIF larger than 5 and was therefore eliminated.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>2.158411</td>
<td>2.244688</td>
<td>1.150063</td>
<td>1.351245</td>
</tr>
</tbody>
</table>

Table 1.3 - VIF of the Second Reduced Model (SRM)

The significances of the variables have also improved.
The normal QQ showed that the points 10, 41, and 43 was outliers and/or influential. The Cook’s distance plot indicated point 10 and 41 specifically as outliers. The dataset revealed no great diverging factor that would class the point as an outlier. The residuals density plot also resembled the normal distribution well which makes the dataset and fit trustworthy when considering the F-statistic. Since the data set was not subjected to any apparent faulty measurements no points were removed.

The stepwise forward elimination as well as the best subset regression algorithms yielded the exact same plots and estimates of the regressor variables indicating that the optimal number of regressor variables were either two or three, i.e. either price ~ income + employment or price ~ income + new_built + employment. In comparison, cross-validation yielded:

Indicating the one variable model: price ~ income
South Sweden

The F-statistic for the full model was large, the p-value below 0.001 and the adjusted-R$^2$ above 0.76. The majority of the regressor variables has a significance lower than 90%, except for the square-meter price which had a significance code of 0 indicating practically no error in estimation. The correlation matrix indicated that a strong relationship existed between the income and square-meter price variables indicating a possibility of multicollinearity.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>newBuilt</th>
<th>household</th>
<th>employment</th>
<th>sqp</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>4.780867</td>
<td>3.131509</td>
<td>1.841993</td>
<td>1.670807</td>
<td>2.352579</td>
<td>5.420288</td>
</tr>
</tbody>
</table>

Table 2.1 VIF of the Full Model (FM)

The square-meter price variable had a VIF larger than 5, the variables was thereby eliminated.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>newBuilt</th>
<th>household</th>
<th>employment</th>
<th>sqp</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>1.909018</td>
<td>2.908184</td>
<td>1.837717</td>
<td>1.649848</td>
<td>2.097887</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2 VIF of the Reduced Model (RM)

For the reduced model all the VIF-values became below 3 and the new significance levels for the current variables is shown below:

The normal QQ and Residuals vs Fitted plots showed that the points 20 and 21 was possible outliers and/or influential. These points was not as prevalent in the Cook’s distance plot and no faulty measurements was found in the dataset so the points will not be removed. The Cook’s distance plot indicated point 43 as a possible outlier, this point was however not present in the other plots and it did
not seem to affect any of the assumptions presented in the theoretical background. The point will thus remain in the model.

The stepwise forward elimination as well as the best subset regression algorithms yielded the exact same plots and estimates of the regressor variables indicating that the optimal number of regressor variables were either two, three or five, i.e. either the present model used, or one of the further reduced models: price ~ income + employment or price ~ income + household + employment.

The cross-validation yielded a suggestion of a three regressor variable model, i.e. the one corresponding to price ~ income + household + employment. The two-variable model was thus discarded and only the three and five variable model were considered in the discussion.

**Småland and the Islands**

The full model gave a large F-statistic and a p-value below 0.001. The adjusted-$R^2$ was above 0.90. The significance of the regressor variables were bad with more than half having a significance lower than 90%, the outliers here being the income variable and square-meter price variable with a significance of 95% and higher. The correlation matrix indicated that a strong relationship existed between the income, household and square-meter price.
Stefan Maraš
Therese Lin

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
<th>sqp</th>
</tr>
</thead>
</table>

Table 3.1 - VIF of the Full Model (FM)

The results from the VIF table above showed three variables with a VIF larger than 5. Since the square-meter price variable had the largest VIF it was removed.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>4.35633</td>
<td>3.60114</td>
<td>1.653400</td>
<td>5.672325</td>
<td>3.249680</td>
</tr>
</tbody>
</table>

Table 3.1 - VIF of the First Reduced Model (FRM)

These results indicate that the household variable had a VIF larger than 5, it was thus decided to exclude the household variable from the model as well.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>1.61343</td>
<td>2.16399</td>
<td>1.109332</td>
<td>1.494649</td>
</tr>
</tbody>
</table>

Table 3.3 - VIF of the Second Reduced Model (SRM)

The updated table indicate that there is little multicollinearity present in the model. The significances of the variables have improved when evaluating the model summary of the SRM compared to the FRM.

These graphs show that 32 was a strong leverage point, that the normal QQ-plot was a bit skewed, and that the residual density plot also seemed to be skewed to the left. The rise of the point was most likely
due to the abnormally large new_built value of 990 which was too high compared to the other values. It thus seems that this value has been incorrectly measured and the decision to remove the point was justified. The results from this removal was a much more balance residuals versus leverage plot indicating a positive effect of the point’s removal. Considering the QQ-plot and the residuals density plot, it would seem that a transformation would had been needed to counteract the skewness. The residuals versus fitted plot did not seem to indicate any characteristic properties of another descriptive function. Since Gaussian assumptions of the residuals were made, the decision to not perform any transformations was justified.

The results of the stepwise backwards elimination unanimously indicated that the final model should only be comprised of one variable. The results of the forward elimination and best subset regression yielded the same results.

The results from the cross-validation also suggested a model comprised of one variable, that variable being the income of the region.

**Stockholm**

The full model had a large F-statistic, the p-value was below 0.001 and the adjusted-R\(^2\) was 0.9. More than half of the regressor variables had a significance lower than 95%. The correlation matrix indicated strong relationship between the income and square-meter price variables.
The income regressor have a VIF above 10 and was eliminated from the model.

All VIF-values were below 3.

The Residuals vs Fitted plot implied that the transformation $x' = 1/x$ would possibly be necessary. The normal QQ and residual density showed that the dataset was close to normally distributed. The density plot resembled the normal distribution well which made the decision to not transform the data justified. Looking at the Cook’s distance plot, point 2 can be identified as an outlier, which was also supported by the normal QQ-plot. However, the values that point 2 represent did not seem to diverge much from the surrounding points, indicating that a measurement error was not present. Since the point did not affect any normality assumptions and he point was chosen to remain in the dataset.
The results from the stepwise backwards elimination were the same as the results from the forwards and best subset selection algorithms. The figure indicates recommendations for three models, those being a two, three, and five regressor variable model.

The results from the cross-validation algorithm implied a model using three variables and so the two-variable model from above is discarded and the final models evaluated in the discussion were thus: consumer closing price ~ household + employment + sqp and consumer closing price ~ rate + new_built + household + employment + sqp

East Middle Sweden

The F-statistic for the full model was fairly large and the p-value was below 0.001. The majority of the regression variables had a significance lower than 90%, except for the square-meter price which have a significance code of 0 indicating practically no error in estimation. The correlation matrix indicated strong relationship between the income, household, and square-meter price variables.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
<th>sqp</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>21.288107</td>
<td>2.725448</td>
<td>1.644203</td>
<td>7.184421</td>
<td>1.837930</td>
<td>29.851679</td>
</tr>
</tbody>
</table>

Table 5.1 - VIF of the Full Model (FM)
The high VIF-value for the square meter price justified the decision to eliminate this variable.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>5.887446</td>
<td>1.916610</td>
<td>1.517881</td>
<td>6.444779</td>
<td>1.559491</td>
</tr>
</tbody>
</table>

Table 5.2 - VIF of the First Reduced Model (FRM)

The income and household variables still had VIF-values above 5. Since the household regressor had the largest VIF, this variable will be removed.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>1.840354</td>
<td>1.916070</td>
<td>1.072556</td>
<td>1.138641</td>
</tr>
</tbody>
</table>

Table 5.3 - VIF of the Second Reduced Model (SRM)

All the VIF-values was below 2 for the second reduced model.

The plots indicated properties of a normal distribution, the residual density plot however seemed to have a tail along its left side which might be an indicator of a needed transformation of the data. The figures indicated that point 16 might be a possible outlier. The original data shows a volatile price decrease compared to point 15 and 17 while having a fairly large new_built value. Since the income follows a reasonable trend this price and new_built value might indicate a faulty measurement. Removing the point yields much more reasonable results with an improved residual density plot and a better Cook’s distance plot as illustrated below.
The greatest difference from the new plots could be seen in the density plot of the residuals which form a much more symmetrically weighted shape than the previous data set yielded. The biggest improvement was seen in the summary of the fitted model where the rate, new_built and employment have all improved their significance level from 90% to 95%. The decision to remove point 16 was thus justified since its origin comes from a measurement error.

The results of the stepwise backwards elimination were the same for the forward and best subset regression algorithms and the unanimous recommendation was for the four regressor variable model presented earlier as the SRM. The cross-validation algorithm on the other hand recommended a further reduced model of two variables: price ~ income + new_built.

North Middle Sweden

The full model had a large F-statistic and a p-value below 0.001. The adjusted-$R^2$ was above 0.90. The significance of the regressor variables however mixed with more than half having a significance lower than 95%, the exception here being the income and sqp variables which had a significance greater than 99%. The correlation matrix indicated that a strong relationship existed between the income, and square-meter price variables indicating a possibility of multicollinearity.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
<th>sqp</th>
</tr>
</thead>
</table>


As shown in the table, the sqp regressor and the income regressor have high VIF-values. The square meter price was first excluded from the model as its VIF-value was the largest.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>3.745663</td>
<td>3.262947</td>
<td>1.597124</td>
<td>3.822959</td>
<td>2.113474</td>
</tr>
</tbody>
</table>

Table 6.2 - VIF of the Reduced Model (RM)

The reduced model gave VIF-values below 4.

From the figures above, it would seem that the points 43, 42 and 26 might have been outliers or leverage points. However, there were no abnormalities found in the dataset concerning these, and the Cook’s distance plot did not yield any great indications of a harmful outlier. No assumptions were impaired by these points.
The results from the stepwise backwards elimination were the same for the stepwise forward and best subset eliminations. The results show a recommendation of two, four, and five regressor variables. Looking at the results from the cross-validation below, a four variable model was the best contender to the RM presented in the VIF-analysis.

The final model between these two is presented in the beginning of the discussion.

Middle Northern Sweden

The F-statistic for the full model was not that large and the p-value well above 0.001. The adjusted-R\(^2\) was 0.74. Only two regressors having a significance higher than 95%. Furthermore, the correlation matrix indicated that a strong relationship existed between the income, and square-meter price variables.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
<th>sqp</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>4.618508</td>
<td>2.514229</td>
<td>1.205806</td>
<td>4.288857</td>
<td>2.075018</td>
<td>5.400194</td>
</tr>
</tbody>
</table>

Table 7.1 - VIF of the Full Model (FM)
The square meter price showed the greatest correlation to other variables as well as a VIF above 5. It was reasonable to exclude this regressor from the final model and thus end up with the reduced model: price ~ income + rate + new_built + household + employment.

The reduced model yields a lower multiple $R^2$ value of 0.75, adjusted $R^2$ value of 0.72 and a little larger p-value below 0.001 indicating that the reduced model might predict the response variable a little poorer than the full model. Calculating the VIFs however yields much better results:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>2.715515</td>
<td>2.494371</td>
<td>1.178665</td>
<td>3.356210</td>
<td>2.072663</td>
</tr>
</tbody>
</table>

Table 7.2 - VIF of the Reduced Model (RM)

This new table indicate that even though the reduced model might have a weaker predictive ability than the previous full model, this model suffers less from the problem of multicollinearity with almost every regressor variable having a lower VIF than their predecessors.

The residuals density nearly perfectly resembled a normal distribution. The Normal QQ-plot however shows some signs of a tail in the upper and lower bounds. This might be an indicator a light-tailed distribution but since the normality assumptions seem to hold, no transformation was performed on this dataset.
The Cook’s distance plot indicates that the original data contained an outlier which corresponded to point 28. The points behavior might be an outlier since the housing price fluctuated up from point 27 and then down to point 29 in volatile manner. However, the most contributing factor must have been the points new_built value which was 279 and exceeds the next-to largest number by 107. The point corresponds to 2012Q1, a year that did not suffer any larger economic difficulties that would justify such a volatile divergence from the remaining data. The points divergence must thus be a consequence of a measuring error when the data was compiled. Point 28 was thus removed from the data set and the reduced model refitted.

The new Cook’s distance plot identified another potential outlier in the form of point 16. Looking at the Normal QQ-plot and the residuals vs leverage plot it did not seem to be of greater importance. Analyzing the data did not show any indication of a measurement error and since the normality assumptions was not impaired the point was justified in remaining in the dataset.
The stepwise forward elimination as well as the best subset regression algorithms yielded the exact same plots and estimates of the regressor variables as the stepwise backwards elimination which indicated that the optimal number of regressor variables were either three or five, i.e. either the RM above, or the further reduced model: price ~ income + new_built + household. The results from the cross-validation below also supported the further reduced model consisting of three variables.

Upper Northern Sweden
The full model gave large F-statistic and a p-value below 0.001. The adjusted-\(R^2\) was above 0.90. Every variable, except of square-meter-price had a significance lower than 90%. The sqp regressor had a significance larger than 99%. The correlation matrix indicated that a strong relationship existed between the income, and square-meter price variables indicating a possibility of multicollinearity.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
<th>sqp</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>7.996691</td>
<td>3.393763</td>
<td>1.217765</td>
<td>4.038337</td>
<td>2.129249</td>
<td>8.947570</td>
</tr>
</tbody>
</table>

Table 8.1 - VIF of the Full Model
The sqp regressor and the income regressor have large VIF-values. The largest variable with the largest VIF, square-meter-price, was excluded.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>income</th>
<th>rate</th>
<th>new_built</th>
<th>household</th>
<th>employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>4.537365</td>
<td>2.571724</td>
<td>1.053600</td>
<td>4.021378</td>
<td>1.460742</td>
</tr>
</tbody>
</table>

Table 8.2 VIF of the Reduced Model

The table above indicates a decrease in the overall VIF-values.

The $R^2$ values have reduced in size to 0.88, the significance of the regressor variables have all improved and the p-value has remained below $2.2 \times 10^{-16}$. An adjusted $R^2$ value of 0.88 was still sufficiently high to indicate a reliable and predictable model, the choice of excluding the sqp regressor was justified.

The residuals versus fitted plot indicated a squared relationship between the $x$- and $y$-plane which needed a transformation. The most suitable transformation to apply here was the $x^2$ transformation since this will counteract the symmetric fluctuation of the residuals vs fitted values. Applying this transformation on the data set yields:

The residuals versus fitted plot indicated a squared relationship between the $x$- and $y$-plane which needed a transformation. The most suitable transformation to apply here was the $x^2$ transformation since this will counteract the symmetric fluctuation of the residuals vs fitted values. Applying this transformation on the data set yields:
The new Residuals vs Fitted plot showed clear improvement from the previous plot. The biggest difference can be seen in the residual density plot which almost perfectly resembles a normal distribution. The normal QQ-plot also indicates a well distributed data set.

The figure indicates that the points 15, 38, and 39 were outliers. These values from corresponding to these values were however not divergent and do not indicate any errors in measurement. Since the normality assumptions still hold as suggested by the Normal QQ-plot and the residuals density plot these points were not removed.
The stepwise forward and best subset selection yielded the same results as the stepwise backwards elimination, i.e. a model containing, two, three, or four variables. The results from the cross-validation algorithm yields an optimal amount of two variables as the graph below indicates:
Discussion

West Sweden

Comparing the three models, these being the three, two, and one variable models respectively, the first one can be rejected since the regressor variable “new_built” does not satisfy the criteria of have a significance of 95% or higher. Comparing the second model to the third model there are some advantages that are clearly present. The model has one more regressor variables while still maintaining a larger adjusted R^2-value indicating that the model is very well fitted. The significance of the regressors and intercept seem to be better for the third model however. This might be a result of the employment variable which has a lower significance code. However, for the purpose of finding the most reliable model possible, it is of interest to choose the model that uses the most information in the most effective way. The final model is for Western Sweden is thus chosen as: **consumer closing price ~ income + employment**

South Sweden

The first thing to point out when comparing these two models is that the adjusted R^2-values are very similar, and their difference is negligible when deciding which model might be the superior predictor. Considering the p-value and significance levels the three regressor model performs much better in this category with better significance levels compared to all the five regressor model’s variables. The F-statistic is also larger and the p-value smaller for the three-variable model. It can thus be concluded that the three regressor model outperforms the five regressor model. The final model for Upper Northland is thus: **consumer closing price ~ income + household + employment**.

Småland and the Islands

There stands a superior final model when comparing the summary of the SRM and the one variable model. The adjusted R^2 value has increased in greater proportion to the SRM and the significance level of the regressor variables have improved. In actuality, the SRM has three variables with a significance lower than 95% while the model as a whole has an F-statistic of 94.6 versus an F-statistic of 405.6 for the one variable model. The smaller model is also absent of multicollinearity since it is only dependent of one variable, making it superior in this aspect as well. In conclusion, the one variable model is, as a whole, a more reliable model to describe the closing prices in Småland and the Islands. The final model is thus: **consumer closing price ~ income**.

Stockholm

Looking at the predictive abilities of the two models in the result, the adjusted R^2-values are practically the same with negligible differences. The biggest difference is found in the significance levels of the different variables. All the regressors in the smaller model have a significance lower than the RM while the smaller model as whole has a larger F-statistic of 1798 compared to the RM’s F-statistic of 1033. This is an indicator that the smaller model is statistically more reliable with less error in its estimates. The smaller model is also cheaper since it has fewer variables, and is thus as whole a more preferred choice of model for describing this region. The final model for Stockholm is thus given by: **consumer closing price ~ household + employment + sqp**.
East Middle Sweden

Comparing the results of the two regressor variable model to the SRM it is noteworthy to see that although the significance has greatly improved for the intercept, the income variable has remained its significance while the significance level of the new_built regressor has increased. Furthermore, the adjusted $R^2$ value has decreased with about 0.01 from the SRM to the two-variable model. In essence, there has occurred a tradeoff between the model’s predictive ability and accuracy, versus the number of variables used. Since the purpose of this report is to explain the regional affordability/purchasing power among consumers as good as possible it is important to use the most accurate model available. For this reason, the choice of the four regressor variable model is justified yielding the final model:

$\text{consumer closing price} \sim \text{income} + \text{rate} + \text{new_built} + \text{employment}$.

North Middle Sweden

As the results clearly show, the new reduced model with four regressors performs very well with great significance, better than the RM presented in the beginning of this regional analysis. The adjusted $R^2$ have also slightly improved while the p-value still stays below $2.2 \times 10^{-16}$. The F-statistic has also improved from 167.8 for the RM to 212 for the four-variable model. This new reduced model, i.e. $\text{consumer closing price} \sim \text{income} + \text{rate} + \text{household} + \text{employment}$ is thus the most superior and the one that will be used to describe the Northern Middle Sweden.

Middle Northern Sweden

Analyzing these results, a conclusion can be drawn that although the VIFs are much better for the three regressor model compared to the RM, the plots and adjusted $R^2$-value are worse as well as having suffering from a lesser significance than the RM. In essence, the larger model trades a higher possibility of multicollinearity (which is fine since its VIFs are not close enough to 5 to be alarming) against the ability to predict this region’s housing prices. It can thus be concluded that the five regressor variable model $\text{consumer closing price} \sim \text{income} + \text{rate} + \text{new_built} + \text{household} + \text{employment}$ is to be used for the Middle Northland region.

Upper Northern Sweden

The results indicate clear differences between the values given by the two respective model summaries. In the smaller reduced model (SRM, two regressor model), the significance level is much better, being measurably lower than that of the RM (consumer closing price ~ income + rate + household + employment). As a matter of fact, two of the variables in the RM have a significance of 90%, well below the cut-off point of 95%. Although the $R^2$-values are lower for the RM, they are so marginally lower that it will not have any remarkable impact on the model’s predictive ability as a whole. It is thus a negligible factor in the holistic comparison that is being made. It is however more interesting to compare the differences in p-value where the SRM has a lower value than the RM with two degrees of magnitude. Furthermore, the SRM has much larger F-statistic of 123.2 compared to the RM’s F-statistic of 53.7. The conclusion that can be drawn is that the SRM model is much simpler and performs statistically better to its RM counterpart. The final model for Upper Northland is thus: $\text{consumer closing price} \sim \text{income} + \text{household}$.
The models

<table>
<thead>
<tr>
<th>Region</th>
<th>Income</th>
<th>Construction</th>
<th>Employment</th>
<th>m² price</th>
<th>Repo rate</th>
<th>Household</th>
</tr>
</thead>
<tbody>
<tr>
<td>West Sweden</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Sweden</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Småland and the Islands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>East Middle Sweden</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>North Middle Sweden</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Middle Northern Sweden</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper Northern Sweden</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

The reason for choosing the macroeconomic factors that are used in the report is based on the beliefs that they have a great effect on the housing market. At first, it would seem counterproductive to remove variables which have a predictive ability of the response variable. However, the desired outcome of the multiple regression analysis is to find the model that best defines the closing price for each region based on a lot of different criteria. In accordance with this, a model that has less variables might have a much better predictive ability and the choice to remove certain regressor variables is thus justified. That one model does not include a specific variable does not mean that the eliminated macroeconomic factor does not affect the housing market. The models may be affected by eliminated variables, however the influence of the other variables are greater, hence the elimination does not affect the models predictability negatively.

One possible reason for this result is that a factor may be eliminated because its correlation to the other factors is too large, which might give rise to the problem of multicollinearity. This means that even though this variable is removed, its effect will still be represented by the remaining variables. Another possible reason is that the variable is very stable and does not fluctuate much in this region, and thus does not affect any changes in the housing market.

The income variable is seen to be present in all the models, except for Stockholm. Disposable income is directly connected to how much money the households have and how much they can spend on investments. This is also the main factor used when calculating the existing affordability indexes HAI and HOI. Looking at Småland and the Islands, the income is the only factor that describes the affordability, which further emphasizes how important this factor is to the households purchasing power.

Employment rate is also a contributing factor in almost every model followed by the change in household. This is not a coincidence either. As more households enter the market and more people are able to afford their own housing property the demand on the market increases which directly affects the closing price of the property.
The square meter price is only present in one of the above models, the one describing affordability in Stockholm. The average square meter price reflects the prices that exist on the market and thereby also the closing price. The correlation is therefore very large among the other regressor variables leading it to the square meter price being eliminated in almost every model.

It is interesting that different regions are affected differently by the chosen variables. It would be reasonable to assume that, since all regions are in the same country, the housing market would follow the same trend throughout the counties. However, it is not too odd that the models are different since the living conditions and priorities are different depending on a consumer’s geographical position. Taking Stockholm as an example, this region encapsulates an area characterized by higher employment rates, supplying jobs that have higher average incomes and finally contributes to the biggest demographic increases in comparison to counties such as Upper Northern Sweden where the farming culture is much more prevalent. The exact reason behind why some factors affect a region more than others is something that could be investigated further in a future report.

The reason for why the different regions are influenced differently by the mentioned, specific macroeconomic factors is not included in this report, however could be an interesting topic of future studies.

Macroeconomic Factors

The present Riksbank’s repo rate is set to a record low of -0.50 percent. This negative rate stimulates the investments and thereby also the housing market. One could consider what would happen when the repo rate leaves -0.50 percent. It will become more expensive for households to lend money for financing housing investments. The house prices may be pushed down as the demand gets lower. When the Riksbank reaches the inflation target it will be natural for the repo rate to rise, this is to prevent inflation to surpass the target. This combined with the fact that we are constructing more houses and that the supply of newly constructed homes on the market is bigger than existing homes may lead to the housing market becoming saturated and the housing prices getting lower. However, this does not necessary mean that affordability will consequently rise since it becomes more expensive to take loans for housing investments (rising mortgage rates).

The effect of inflation is also needed to be taken into consideration for other variables as well. Since the income is increasing it could be premature to think that a larger disposable income would lead to higher affordability among households, i.e. greater purchasing power among consumers. However, this may not necessarily be the case. If the inflation is higher than the increase in wages the affordability may sink. Only when the increase in wages is larger than the inflation does the real impact of the wage growth show. Which could give rise to higher affordability among Swedish households. However, it is important to remember that higher wages also lead to a higher inflation. If the average wage level becomes higher than this may drive the housing prices up, in effect, lowering the consumers’ purchasing power.

Higher employment rates will also affect the affordability since it could increase the demand on the housing market. More people will have money to spend and thereby have the possibility to choose among different properties instead of taking the cheapest on the market. Also, the positive increase of households will boost the demand on the housing market, since there are more consumers present. This will affect the prices and the affordability among customers. Without new construction it could mean that the average prices raise.
Purchasing power
For this report, the closing price was chosen as the response variable. The reason is simple, the closing price shows what the household are able to afford and purchase for as it is the final price settled upon between the seller and buyer. However, there are some problems one must take into consideration. The households may not be purchasing for the maximum level that they can afford, leaving money for other future investments. One household may also buy several housing properties leading to the response variable to be misleading since one household may afford more than just one home.

The HOI and the HOX indexes are both based on income as the main factor for measuring affordability. These values are widely used around the world and are approved by many economists studying the housing market. It is therefore no coincidence that this report’s results also show that income has a big influence on a household’s ability to purchase property on the housing market in Sweden.

Data
Some problems were also faced when collecting data for the regressor variables. The population growth and disposable income only had data gathered by year. To convert the yearly data for income to quarterly data several methods were considered. One way to do this is to set the four quarters in one year to the same income, for example to the average income for that same year. Another way is to use linear interpolation. Since the wages are continuously getting higher each year (approximately following a straight upward line), linear interpolation was the best and most straightforward method to construct quarterly data.

However, one must also consider that factors such as income and population growth do not have to change the same amount every three months. This can be seen by observing the data of the demographic change for the quarters in 2016 and 2017. Many other factors, such as inflation or weather, affect the development and contribute to making the values for the quarter to differ from one another. Therefore, the methods for constructing the data for income and household development can be questioned. Other methods than linear interpolation could be used to get more reliable data for each quarter. However, the aim of this report is to construct models for the affordability on the housing market and not predicting the changes for income, and therefore the most straightforward method was used. This might affect the reliability of the model negatively but is in this case negligible since the results converge with pre-existing theory as previously mentioned.

Future analysis
There are many possible areas to investigate for further analysis of the purchasing power on the Swedish housing market using this report as a guideline. There are also many ways to improve the models to achieve a more trustworthy result.

One possible way is to include more regressor variables. It is obvious that the housing market is affected by more than six factors, and that it is not only macroeconomic factors. By adding more factors in consideration when constructing the regression model, a better reflection of reality could be achieved. However, too many variables with few data points may cause multicollinearity as described in previous sections. Other factors such as inflation (CPI or CPIF), GDP, age distribution or mortgage rate could be included in the model as possible regressors. One highly interesting factor that could be
included is the amortization requirements. The amortization requirements in Sweden have been changing during the last couple of years affecting the housing prices and the households purchasing power.

More data points give more precision. The data used for our models are continuously updated by SCB and Macrobond. In this report only 43 data points were used for each variable, from the second quarter of 2005 to the fourth quarter of 2015. When new data is out, the data points can be added to the excel-files and thereby concluded in the models to get better equations. This will give better estimation of the future purchasing power. If possible, to further improve the fit of the models, monthly or daily data division can be used in the regression model instead of the previously used quarterly division. This can however be a complex data structure to use since data points have to exist for each of the regressor variables, something that SCB or Macrobond does not supply for all of the variables. One possible solution would be to use a separate regression model for each regressor variable to fill in the missing data points. This is however a superfluous solution since the given results have been satisfactory in accordance with the purpose of this report.

In the report, multiple linear regression analysis is used to describe the market. However, the housing market does not necessary follow a linear line, at least, by observing the data, the housing prices do not. It is therefore possible to use other methods, non-linear regression, to construct a new model using the same data. This could yield more precision and a better fit.
Conclusion

There are several macroeconomic factors that affect the affordability on the Swedish housing market. Using multiple linear regression analysis eight separate models were created to describe the affordability in each region. The results show that the purchasing power, in the different Swedish regions, is affected the most by the factors presented below:

**West Sweden:** Income + New construction + Employment  
**South Sweden:** Income + Household + Employment  
**Småland and the Islands:** Income  
**Stockholm:** Household + Employment + Square meter price  
**East Middle Sweden:** Income + repo rate + new construction + employment  
**North Middle Sweden:** Income + repo rate + household + employment  
**Middle Northern Sweden:** Income + repo rate + new construction + household + employment  
**Upper Northern Sweden:** Income + Household

The disposable income is found in seven out of the eight models. Changes in household and the employment rate are also found in the majority of the models. These three thus become the macroeconomic factors that have the biggest influence on the consumer’s power to purchase on the housing market in Sweden. Square meter price was eliminated from all regions except for Stockholm, which makes it the most insignificant among the chosen factors when looking at the country as a whole.
References


Fodérus, Leo; teacher at the department of industrial engineering and management, Kungliga Tekniska Högskolan. Stockholm, lecture 5, 2018-01-30


Macrobond database, 2018


Nordea Markets database, 2018


Statistiska Centralbyrån (SCB), 2018


