May 2013

House prices in Denmark
An in-depth analysis of the price movement

Authors: Duy Quang Do & Karl Gustav Mattias Holten
Supervisor: Ana Alina Tudoran

DEPARTMENT OF BUSINESS ADMINISTRATION
Aarhus University
Business and Social Science
2013
Abstract

We examine the Danish house price index by using time series analysis. We choose a combination of an autoregressive integrated moving average (ARIMA) model and a linear regression model. Eleven independent variables were selected for the model based on past literature and empirical evidence. Using the ARIMA model, we found that the Danish house price index is affected by the three previous quarters. After testing, two models were selected, model 1 and model 2, by using statistic criterion. Model 1 contains all elevens independent variables with two control variables and the three lagged values of the house price index. This model, statistically, has one significant independent variable (consumer confidence index) and has an adjusted R-squared of 0.8823; this means that 88.23% of the variations in the Danish house price index can be explained by this model. Model 2 was optimized to fit a parsimonious model. Four independent variables are proved to be significant in this model (consumer confidence index, population, consumer price index and interest rate). This model has an adjusted R-squared of 0.9087; therefore, 90.87% of the variations in the Danish house price index can be explained by this model. Consumer confidence index proved to the most significant variable in both model. This is a new finding since this variable has not tested before. The ARIMA model had 79.1 % correct when an in-sample forecasting was conducted, which was the best result when compared to model 1 and model 2, 58.3 % and 45.8 % respectively.

Acknowledgement

We would like to thank our associate professor, Ana Alina Tudoran for her comments and guidance. We would also like to thank Martin Christensen from Green Analysenstitut for giving us access to the data of their consumer confidence survey.
# Table of Contents

1. **Introduction** .............................................................................................................................. 5  
   1.1 Structure of the thesis .................................................................................................................. 6  
   1.2 Delimitation .................................................................................................................................. 7  
   1.3 Technical tool ............................................................................................................................... 8  

2. **Literature review** .......................................................................................................................... 9  

3. **Trends in Danish housing market** .............................................................................................. 11  
   3.1 First cycle ...................................................................................................................................... 11  
   3.2 Second cycle ................................................................................................................................ 11  
   3.3 Third cycle .................................................................................................................................... 12  
      3.3.1 Strong fundamentals .................................................................................................................... 12  
      3.3.2 The bubble ................................................................................................................................. 13  
      3.3.4 Bubble burst ............................................................................................................................... 14  
   3.4 Some Remarks ............................................................................................................................... 14  

4. **Theoretical background** .............................................................................................................. 16  
   4.1 Dependent variable ....................................................................................................................... 16  
   4.2 Independent variables ................................................................................................................... 17  

5. **Data collecting and sampling** .................................................................................................... 24  
   5.1 Dependent variable ....................................................................................................................... 24  
   5.2 Independent variables ................................................................................................................... 24  
   5.1 Data Adjusted .............................................................................................................................. 26  

6. **Hypothesis treatment** ................................................................................................................... 27  

7. **Methodology** .............................................................................................................................. 29  
   7.1 Decomposing a time series ............................................................................................................ 29  
   7.2 Unit root test ............................................................................................................................... 31
7.3 Difference the equation ........................................................................................................33
7.4 Autoregressive integrated moving average model ..................................................................34
7.5 Combination of linear time series model ..............................................................................34
7.6 Model selection criteria .........................................................................................................36
7.7 White noise test ......................................................................................................................38
7.8 Forecasting ............................................................................................................................39

8. Analysis and Result ..................................................................................................................40
8.1 ARIMA model: .......................................................................................................................40
8.2 Combination linear regression model ....................................................................................49
  8.2.1 The model 1 .....................................................................................................................52
  8.2.2 The parsimonious model ...............................................................................................54
8.3 Forecasting ............................................................................................................................58

9. Conclusion ................................................................................................................................64
  9.2 Further improvement ............................................................................................................66

References ..................................................................................................................................67

Appendix
1. Introduction

During the burst of the U.S housing bubble in 2007, Treasury Secretary Henry Paulson said that the problems in the housing market, particularly, the housing prices represented “the most significant current risk” to the US economy (AFP, 2007). Denmark was no exception, in the same year, the Danish housing bubble busted leading to significant falls in housing prices pluming the entire Danish economy (Danske Bank, 2008).

The developments in the Danish housing market over the last couple of years from the long boom in prices that started in 1992 until the turbulence in the recent years have shown how big of an effect the housing market has on the entire Danish economy (Danske Bank, 2008). The movements in the housing market affect the entire economy such as authorities, banks, home owners, non-home owners, investors and other economic agents. The housing bubbles followed by the bursts have proven to have a devastating effect on the economy of a country (Thornton, 2006). The house price has become a central economic index for society; the growing relevance can be seen in the extent of newspaper articles on the subject. Statistics Denmark began in 2006 to calculate the housing price index on a monthly basis instead of quarterly as before.

The price of a house sold is likely to be influenced by the previously sold houses hence; the data is not based on independent observations. Therefore, it is not appropriate to use univariate statistical methods based on independence such as bar chart or frequency distribution (Woodward et al, 2012). Thus, in this paper, time series regression will be used to analyse the housing price in Denmark.

Applying time series to modelling housing price dynamics is a traditional approach (Gourieroux, 1997). Time series approaches are generally motivated by the presumptions that there is a correlation between the adjusted points or observations in time. This can be explained in term of the current value and values of the past. Time series creates a parametric function of the current value and the past values to focus on modelling some of the future value. This model leads to the use of time domain as a forecasting tool which is very useful for economists because by forecasting the future, it is possible to predict the evolution of prices and estimate the demand and possible effects on the overall economy (Woodward et al, 2012).
As mentioned before, the focus of this paper is on Danish market and creating an adequate forecasting model for the Danish house price index could provide useful information for banks, economic agents, authorities, consumers and a relevant extension in the research field.

The objective of this thesis is to make an in-depth analysis of the Danish house price index in the last two decades. All variables found relevant in the previous literature and research will be combined in a model and tested towards their relevance. The previous studies on the house price index often focus solely on economic fundamentals (e.g. unemployment, income) and their effects on the house price (Wagner, 2005; Skaarup & Bødker, 2010). The effect of behavioral finance (individuals and institutions economic behavior) has become a highlighted issue on the house market in the recent years (Case & Shiller, 2004; Shiller, 2008; Lunde, 2008). This thesis will try to include the markets behavioral aspects and combine them with fundamental economic variables towards the house price index. Based on the above mentioned, the main purpose of the thesis is to create a model that can explain the movements in the Danish house price index. Thus, the following research questions are generated to support this purpose:

- *Is the consumer behavior research significantly relevant towards the house price?*
- *How accurate can the model produce a feasible forecasting outcome?*

As a sub quest to the overall goal of creating a model that explains the Denmark’s house price index an attempt to forecast the house price index will be carried out provided a successful outcome of the main objective.

1.1 Structure of the thesis

This thesis starts out by reviewing some important previous literature regarding the house price dynamics in “Literature review”. Section 3 introduces the previous trends in the Danish housing market, from the 1970’s up to the time period that the data is collected. In the end of this section, the findings and eventual lessons to be drawn from the historical price index will be discussed. Section 4 presents the theoretical background on the house price index and discusses the relevance of the variables selected for the regression. In section 5, the practical presentation of the data and the collection will be made. Section 6 will present the hypotheses and the general model of this thesis. In section 7, different methods and theories will be discussed in order to find the right
parameters for the general model in section 6. In section 8, statistical test will be carried out. Applying the theories found in section 7 to find a feasible that matches the hypotheses in section 6. And the last section, section 9 will conclude what the paper has done and discuss further improvements for future research.

1.2 Delimitation

Much of the recent research on Danish house market has focused on the recent housing bubble. This thesis does not attempt to give an in-depth analysis of the recent bubble or the effects it has on the economy. However, the housing bubble occurred under the period that will be looked into hence it cannot be overseen, as for possible effects it might have had on the housing price.

The overall index is too broad to use in defining the housing price, because by including different kinds of properties, it might affect the outcome of generating the model and provide an infeasible outcome with unclear patterns. For example, agricultural properties might have variables (quality on the dirt) affecting the price of the property that is not relevant for prices of apartments, or houses. Therefore, this thesis will look only at the price index of the market for one family houses. One family houses is basically the normal real estate for the end consumers (first time buyers, families, single persons, senior persons, etc). The focus of this thesis will be on the Danish house market as a whole. A more extensive discussion and validation on the choice of price index will be described further in section 4.

In order to perform a valid time series analysis, it is essential to have a reasonable delimitation of the research area. The data collected for the house price index is taken from observations from 1992 to 2012. In this time period it is possible to obtain quarterly data of the house price index. The collected independent variables which are going to be tested against the house price index are all based upon the findings in the previous studies. This process is further described in detail section 4. The time span may in some cases vary between different independent variables due to the availability of data and observations from valid sources. The process of data collection and conversions is described in further in section 5.

The time series regression assumes that all the independent variables follow a linear relationship therefore, the nonlinear relationship will not be taken into account.
The data was collected from two statistical databases i.e. “Statistics Denmark” and “Danmarks Nationalbank”. Both of these are official and highly trusted databases that offer valid and reliable data.

Time series analysis methods are often classified into two major categories: Time domain and frequency domain methods. The time domain method includes the analysis of the correlation structure, and then developed models will describe and forecast the future behavior in which data evolves in time. Frequency domain method is developed to understand the time series data by examining the data in the perspective of frequency content using different analysis tools such as power spectrum (Woodward, 2012). However, this thesis will mainly focus on the time domain method.

1.3 Technical tool

R is open source statistical software that has been incorporated many years in researching statistical and numerical computing (McLeod, 2011). R is a high quality software that runs on all common computer platforms such as Linux, Windows and Macintosh. Unlike SPSS which provides point-and-click and graphical-user interfaces, R is a command-oriented that users have to type commands and then R will interpret responses interactively to these commands (Fox, 2005). Using a command-oriented interface statistical software provides a few advantages as it is easier to correct, modify and replicate the analyses and each R command and output can be accompanied by an explanatory text (Fox, 2005). Last but not least, R is free and easy to access from anywhere with an excellent build-in help system for beginners (Verzani, 2002) Hence, R project will be used to analyse the time series regression throughout this paper.
2. Literature review

The aim of this section is to make a brief overview of previous literature on house price that is important to this paper. Some of these articles study the housing price at a national level in Denmark while others do international research and comparisons across countries. They use different methods and approaches but they all seem to agree on the relevance and importance of certain factors and variables. The process of selecting the right variables and collecting the data is described later in detail in section 4 and 6 respectively.

“Structural factors in the EU housing market” is a heavy and extensive report from the European Central Bank in 2003. The report presents cross country information and reviews an extensive amount of factors and explains their relevance in housing price dynamics (e.g. housing tax/subsidy, rent dynamics, mortgage market). The report is divided into sections of different aspects of the housing market, but “House price dynamics and their determinants” is the most relevant part for this thesis. Furthermore, the report presents an overview of the development in the housing markets since the 1980’s and explores the origins of the housing price fluctuation (micro/macroeconomic factors). With a large amount of data provided that supports the report’s different statements, it can be seen as a dictionary for housing price dynamics. The report states that house price fluctuations are most often explained by economic fundamentals. These fundamentals (e.g. interest rate, household income, unemployment) are examined in-depth, listed and ranked towards their relevance on house prices.

Girouard, N. et al. (2006) explore the fundamentals behind the house prices in the OECD countries. This report examines the development in the house market over the past 35 years. They especially focus on how to explain and interpret the long lasting increase in house price that took place in the OECD countries since the mid 1990’s up until 2006. The underlying determinants of house price are established by gathering a range of empirical studies and literature that use econometric models to test their relevance. These fundamentals are for example income, interest rate, demographic development and they further emphasize that the housing price in Denmark and some other OECD countries are broadly in line with these determinants.
In 2010 Skaarup & Bødker wrote a paper on the behalf of the Danish Ministry of Finance called “House prices in Denmark: Are they far from equilibrium”. They went through the past trends in the housing market and derived a traditional demand-supply housing model to examine the state of the Danish housing prices. The variables they chose for their model are income, interest rates, construction cost etc. These series are put into time series and tested for their trend, stationary and the order of integration before creating a model of VAR as known as Vector Autoregressive Regression. They concluded that the price increased up until the housing bubble was supported by economic fundamentals: increasing incomes, falling interest rates etc. As for the price after 2004 (it was not possible to explain solely by these fundamentals therefore) they introduce the thought of irrational financial behavior or as they phrase it: “increasing risk appetite by households and financial institutions”.

What all the previous studies have in common is that they explore the dynamics behind the housing price. The models they build are focused solely on explaining the variation behind the house price with no intent to forecast it. However, one of the more recent studies that take on the challenges of forecasting the house price index is Bork & Møller (2012). Their approach is to include a large number of economic time series to conduct the testing. The three factor model that they create consisting of 122 different time series on variables manages to explain about 50% of the variation in house price movement one quarter ahead. They conclude that the three factor model that they produce performs well in comparison to autoregressive benchmarks with lag structure and computational intensive factor forecast combination models.

These four studies all together cover the subject of house prices from several different angles. By exploiting the knowledge and the findings in these papers, it is possible to start selecting variables for a housing price model.
3. Trends in Danish housing market

During the past four decades (since the 1970), the Danish housing market has experienced three major price cycles with large increases followed by decreases in the house price index (Skaarup & Bødker, 2010). It is in the interest of this thesis to include a short review of the first two cycles and a more in-depth and detailed review of the last one. The last cycle is especially important because it falls under the same time span as the testing of the house price that is conducted later in this thesis. The explanation of the price movements in these cycles does not always lie within the factors that are normally considered to have a relevant effect on the house price. Looking at the historical trends will create a greater understanding on the complex house price index. The final findings might prove to be useful to consider when conducting research on the mechanism behind the house price index.

3.1 First cycle

The first cycle started in the beginning of the 1970’s and was started shortly after a small decline during the first oil crisis in 1973. There was a real price increase by 25,6% from 1974 up until the second oil crisis in 1979 (Skaarup & Bødker, 2010). The “shock” increase in the oil price had a great negative effect on fundamental factors behind the house price because the unemployment rate rose, household income fell and the interest rates went up (Det økonomiske råd, 1979). The housing price fell with a total 36,8% in real terms in the next three years making it the largest price fall since the Second World War (Realkredit Denmark, 2012).

3.2 Second cycle

The second cycle began in 1982 with increasing prices lasting until 1986. As an effect of “Kartoffelkuren” launched in 1986, prices peaked and fell straight for 29 quarters until 1993 with a total price fall of 36,4% in real terms (Realkredit Denmark, 2012). “Kartoffelkuren” (“the potato package” initiative) was a finance-political initiative launched by the Schlüter-government. The goal was to reduce the Denmark’s trade deficit and encourage savings. Consumer and house loans were through legislation made more expensive and harder to obtain (Gyldendal, 2009). The construction of new houses fell drastically. The year after, in 1987, a tax reform leading to higher taxes boosted this negative development for house prices. (Wagner, 2005)
3.3 Third cycle

Figure 1: Price index: One Family Houses

The last major price cycle beginning in 1993 is the most relevant as it overlaps the research period of this paper on the Danish house price index. According to Skaarup & Bødker (2010) this cycle is unique in its length and real price increase. As mentioned in section 3.1 and 3.2 the length of the upturns and downturns of the previous cycles was between three to five years (13-18 quarters) while the upturn of this cycle lasted for 14 years (55 quarters) between 1993 and 2007 causing a stunning increase of 176.6% in real price (Skaarup & Bødker, 2010).

The cycle period between 1993 can with be split into 4 periods with different variables and events (strong fundamentals, the bubble, the bubble burst) driving and influencing the house price index.

3.3.1 Strong fundamentals

When the upturn started in 1993, the housing market was coming out from a low point, being "depressed" and low valued. From 1993 to the first quarter of 2004, there was strong economic fundamentals; the rate of unemployment was low, the real interest rate decreased, tax froze, the nominal income increased behind the increasing house prices (Skaarup & Bødker, 2010). Wagner (2005) stated that about 90% of the price increase in the house price index could be explained by these fundamentals. The strong development on the housing market was further supported by the new legislation which allowed the introduction of interest-only loans, fixed interest rate loans and adjustable interest rate loans on mortgage market in 2003 (IMF, 2007)(Lunde, 2008). The year after
in 2004, mortgage loans with a guaranteed interest rate was introduced, adding more options and products to the Danish mortgage market (IMF, 2007).

3.3.2 The bubble

In the period between the first quarter of 2004 and the third quarter of 2007, the housing price index increased by over 67% in nominal terms (appendix A). The house price index reached the highest in the Danish history and was described as being on an exceptionally high level (Lunde, 2009). This rapid price increase in housing prices and the consequences for the whole economy in the aftermath of the bubble burst matches Mark Thornton’s description of the U.S housing bubble in “The Economics of Housing Bubbles”. Skaarup & Bødker (2010) state that the introduction of new loan products increased the options for people to borrow combined with the strong economic fundamentals caused the bubble. Evidently in July 2007, the interest-only loans that had been introduced in 2003 accounted for 42.3% of all outstanding mortgage for owner-occupiers (Danmarks Nationalbank, 2007). Økonomi og Ervervsministeriet supported the view that interest-only mortgages were a strong driver behind the house price development in paper published 16/9 2008. The question of behavioral finance (the psychological aspect), people’s economical behavior was raised before and especially in the aftermath of the bubble. Research has been done in this field in connection with the U.S housing bubble and due to the similarities between the two bubbles, the findings are interesting from a Danish perspective (Value walk, 2012). These theories might help explain the anomalies that occurred in the market during this period and cannot be explained by the traditional economical theory.

“A tendency to view housing as an investment is a defining characteristic of a housing bubble”.

(Case & Shiller, 2003, p.321)

Robert J. Shiller (2008) points out the overconfidence in the housing market as a crucial factor towards the housing bubble. Believing in a never-ending increasing market, buyers believed that investment in housing was good at any given time or price. For buyers, housing now was not only bought with the purpose of occupancy but as a mean to acquire wealth (Mortensens & Seabrook, 2009). The high return on equity for house investments at the time due to the rapid price increase led the speculation to boom even further in the housing market (Lunde, 2007).
The price race left the Danish house owners with the biggest mortgage debt compared to disposable income out of all OECD countries. Already in 2005, during the bubble, it was at a rate of 260% (Girouard et al., 2008).

### 3.3.4 Bubble burst

The house market peaked in 2007 when the bubble busted and caused a large decline in overall housing prices. Between the period from the last quarter of 2007 to the last quarter of 2009, prices fell by 18% in nominal terms (appendix A). Why the bubble busted at exactly this point has yet been determined as Lunde (2008) argued that there was no specific “shock” or “trigger factor” that started this fall. Possible causes might have been the U.S financial crisis striking in 2008 and the fact that the Danish economy had a decline by 1.3% in the GDP sending Denmark to its first recession since 1994 (Danmarks Nationalbank, 2009). The initial decrease was followed by a slight nominal increase of 3.4% in the following four quarters. Between the third quarter of 2010 and up until the latest available data of the third quarter of 2012 the house price set a stable total decrease of 7% in the nominal terms.

### 3.4 Some Remarks

The house price is normally derived from fundamental economic factors that will be selected and described in section 4. These are factors that have been proven to have a relevant effect on the house prices. There is no evidence from the past 40 years in the trends that oppose their relevance. However, by studying the trend of the past and the three price cycles they all seem to have one thing in common. Each of them has factors, normally not considered relevant for house price models which have come to dominate and trigger these cycles. One might call them the “jokers”. For each of these cycles, the dominant or triggering factor has been different. For the first cycle in the 1970’s, this factor was the oil price. The second cycle in the 80’s, it was the finance political legislation and decision making that changed the cause of the house price. In the third and the current cycle, the mortgage deregulation and the consumer’s financial behavior were behind the bubble that this cycle is famous for. While the behavioral aspect of consumers purchasing decisions is likely to always have had an impact on the housing market, it was not until the last cycle that it became a highlighted and ”relevant” factor in the academic world. One rising question and concern is how to treat these variables when doing up to date research on the house price index. There is little or almost no evidence that the oil price has a significant or direct effect towards the house
price today. The finance political decisions and the mortgage legislation undisputedly have had a large effect on the market before and will have so in the future. The potential of using these factors in a pricing model or to forecast the future price is very small. The new legislation can be presented at any time and offers no tangible data or numbers to be placed in a model, the outcome and effect on the house price is hard to predict for anyone, even the legislators. As for the behavioral economic factor, this paper has been given an opportunity to test its relevance towards the house price index. By being given the access to a unique research containing the data on the Danish people’s expectations towards the housing market, this variable will be presented in the section 4.2.
4. Theoretical background

Looking at the past literature, research and evidence, the role of economic fundamentals is undoubtedly always present as strong and consistent determinants of the housing market. The house price can be said to be determined by the demand and supply (Kim & Renaud, 2009). In this section, the dependent and independent variables chosen for this thesis are presented and argued for. These variables will be used in the time series model to test against the house price.

4.1 Dependent variable

As stated in section 1.2, the house price index for one-family houses will be used for the analysis:

*House price index; One family houses*

Price indexes are offered for a vast range of different kinds of property (e.g. one-family houses, residential properties, building sites, and owner occupied flats).

This kind of property is commonly denoted as just “houses” or “residential homes”. In 2008, 50% of all Danish families owned one family houses (Lunde, 2008).

Figure 2: Sales of real property in Denmark

Source: Statistics Denmark

As can be seen from figure 2 extracted from Statistics Denmark with quarterly data from 2006 to 2012, the sales of one-family houses dominates the market and is by far the most common transaction. One family houses are also by far the most common type of property in Denmark (Mattsson, 2006).
One family houses is the property type which is frequently mentioned in newspapers and media. The developments in the HPI of this category is followed closely by the government, media, economists and if not least the house owners themselves. The one family house price index is frequently used as an indicator for the whole Danish house market (Danmarks Nationalbank, 2009, Lunde, 2008 and Lunde, 2009).

There are regional differences when it comes to the house price in Denmark. In the metropolitan area such as Copenhagen the prices has historically been higher and bared a higher volatility than on a national level (Skaarup & Bødker, 2010). The regional price differences are often simply explained by the stronger economic fundamentals behind regions with a general higher price such as lower unemployment or higher nominal income (Skaarup & Bødker, 2010). Breaking the house price index down into regional levels may be an advantage because it can provide a more detailed view on specific areas of Denmark. However, the regional perspective of the housing price might be interesting, gathering regional data on the house price would require that all the independent variables relevant to the house price should be collected at regional levels (Manning, 1995). Data like this could be very difficult or impossible to obtain. This would require more time, resources and space than this thesis has to offer or a few regions could be chosen in the sacrifice of the others. The regional differences in price index for one family houses should on the other hand not be exaggerated. They follow the same distribution pattern and movements, upturns and downturns. In the case of owner occupied apartments for example, the regional differences could play a vital role since most apartments are allocated around the capital area and big cities resulting in a very uneven and uncorrelated price movement compared to other regions.

The goal of this thesis is to gain an overall perspective of the whole Danish housing market; hence it is both irrelevant and unnecessary to divide the house market into regional areas.

4.2 Independent variables

This section is a vital part of this thesis in choosing which variables to include in the model for the Danish house prices. It will determine the outcome, reliability and validity of the model. The variables for this thesis have been carefully chosen based on a set of theoretical and practical criterion. For a variable to be included in the analysis it has to be proven as a significant factor by the previous research in the field of house market. This could be defined as acknowledge
publications of research from governments, universities, organizations & institutions and different economical agents on the house market. Having passed the theoretical criteria, some practical issues arise. The practical outcome and the presentation can be found in section 5. At first the data on the variable has to be available. It has to be frequent and able to be obtained within the required time period.

There are many different variables that have small as well as big effects on the housing market. The independent variables gathered in this thesis are the variables that have been proven to have significant impacts on the housing price by previous studies and research. The independent variables used in this thesis are described as follows:

**Unemployment and total employment**

The labor market is a key ingredient in the fundamental macroeconomics that is commonly used to explain the house price movements. It is often used in studies and other research on the house price (Girouard, N. et al. 2006, Case & Shiller 2004 and Lunde 2008). The statistics of the labor market can be presented in many forms but for a national perspective, two variables are normally used to describe the status of the market, the data on unemployment and on the employment.

Skaarup & Bødker (2010) use unemployment variable in their paper to allocate the equilibrium of the house price. They find it significant and negatively correlating towards the house price. An increasing unemployment rate will lead to a decreasing house price index and vice versa. In periods of very low unemployment rate, it may lead to less a risk-aversion among households and institutions. With a sense of a great job security during such periods the unemployment rate might spark the house price above the normal correlation rate (Skaarup & Bødker, 2010).

Numbers on employment is another statistics describing the status on the labor market. Although they might seem to represent the same thing, e.g. subtracting the unemployment rate from the total available workforce does not equal the total amount of employed people, both statistics are subject to different standards and face different issues (Hussmanns, 2007).

Bork & Møller (2012) conducted a research on housing price and found that the employment variable is a significant factor when forecasting the US house price. Furthermore, Case & Shiller (2004) also included the employment variable in their model to see if there was a bubble in the U.S housing market. Including both unemployment and employment variables to the model excludes all
the debate and issues between which of the two variables is better, securing the capture of the labor markets relevance towards the house price.

Population

Demographic variables and changes are frequently and reoccurring in the housing price literature and broadly viewed as a fundamental and long term determinant of the housing price. The long term population trends and especially an increasing long term population trend is a part of a “scarcity land theory”. Land is a scarce resource and with an increasing population there will eventually be an effect on the house price as demand for housing increase (ECB, 2003). Demographic changes are believed to hold relevance towards the house price as it affects the demand of housing (Skaarup & Bødker, 2010). There are some other examples of papers that also use and argue population as a fundamental variable for housing price analysis such as Case & Shiller, 2004 and Girouard, N. et al., 2006. A positive correlation is expected as an increased population would increase the demand for houses, especially in a long term perspective. In a recent publication David Miles (2012) suggests that the evolution of the population density is a major determinant for house prices rises next to income. The predicted increase in the population density is believed to present great challenges to the U.K housing market (Miles, 2012). Statistics Denmark predicts that the Danish population will increase from 5 580 516 people in 2012 to 6 158 633 in 2050 (DST, 2013), an increase of 578 117 people. Based on previous research and findings on the population, there is a strong evidence supporting population’s relevance towards the house price index, hence it will be included in the model.

Land price

The relationship between the house price and the price of land is very logical and straightforward. A higher price on land increases the cost of building new houses which lead to a higher overall price for houses. The “scarcity of land” is a concept or theory that is regularly appearing in the housing market literature. Due to its nature as a limited resource, the availability or the supply of land suitable for housing construction declines over time, the price will increase in the long run (ECB, 2003). Evidently, this will drive the housing prices up. Wagner (2005) uses the scarcity of land as a possible explanation on the vast increasing house prices in the Copenhagen capital area during the period from 1993 to 2005. During their review of the past literature Girouard, N. et al.
(2006) present three empirical studies on house price determination that have found land price to be important in Denmark (Wagner, 2005), Ireland (McQuinn, 2004) and Japan (Nagahata et al, 2004).

ECB (2003) argues that in the absence of reliable data, the price of land combined with the construction cost based on a weighted average of land price (30%) and construction cost (70%) to determine the house price. Bourassa, N et al (2010) concluded in their research of the price of one family houses in Switzerland during the period of 1978 to 2008 that:

”Over the course of a property cycle, house price changes are largely driven by land price changes”

(Bourassa, N et al, 2010, p.25)

Interest rate

Interest rate is one of the most powerful fundamental determinants of the house price (ECB, 2003). When people buy a house, the most common way of financing this investment is through a mortgage loan. Currently consumers can borrow up to a maximum 80% of the value of their one family houses in Denmark (Totalkredit, 2013). The interest rate determines the cost of borrowing this money. Hence the movements in the interest rate will increase or decrease the demand for housing as borrowing money becomes cheaper or more expensive. In 2007, the total outstanding mortgage debt for Danish households were 1 935\(^1\) bn.kr or 114% of Denmark’s total GDP that year of 1 695,3bn.kr\(^2\). The interest rate for the consumers is individually set by the bank depending on their risk evaluation, equity and collateral etc. The biggest determinant of this rate though, is ultimately set by the National Banks lending rate which is their rate for lending money to banks with collateral as security. The changes in the interest rate by the National Bank drag down the economy, through the mortgage institutions and finally to the end costumers, the households. (Danmarks Nationalbank, 2008).

In their VAR model Skaarup & Bødker (2012) found out that the interest rate is semi-elastic to the house price. A one percent point decrease in the interest rate would lead to about 18% increase in the house price. International studies like Girouard et. al., 2006 found the same but much smaller negative correlation between the interest rate and the housing price.

\(^1\) Danmarks Nationalbank 2007
\(^2\) Statistics Denmark, 2010
**GDP Growth change**

The gross domestic product is the sum of all products and services produced and consumed or in other words, the overall economy of a country for a certain period of time. The change between two periods reveals if the economy is shrinking or growing. Lunde (2008) discusses the role of recession in the overall economy for the housing bubble, using GDP growth as an indicator. As GDP is a measurement of all the economic activities in a country, it is not a fundamental factor itself, however it is included to test how the movements in the overall economy affect the house price index. Gregory Sutton (2003) finds that house prices correlate positively towards the growth in GDP a 1% increase in the latter resulting in a 1-4% increase in real house prices.

**Construction cost**

The construction cost variable contains the total cost of labor and material for house building and it is an important variable to include in a house price model. It is natural to assume that the house price is positively correlated with the construction cost (Skaarup & Bødker, 2010). If the price of constructing new homes increases the housing price in general will increase. In the lack of reliable housing data, the ECB (2003) used estimations from Bundesbank combining land price and construction cost to estimate house prices were the construction cost stood for 70% of the house value.

Tobin q is a finance investing theory and is used in housing to describe the correlation of the housing price and the construction cost. It is defined as \( \frac{P}{C} \) and suggest that there is a long run equilibrium in the market when the price of existing houses, \( P \), is equal to the construction cost plus with the land price of new homes, \( C \), or a Tobin q of 1 (Haagerup, 2009). If Tobin’s q is above 1 there is a positive profit to gain from constructing houses as the house price exceeds the construction cost and vice versa, hence it is determinate for house investments. Tobin’q is a general theory and will not be taken into account in this thesis, however it is an indication on the importance of the correlation between the housing price and the construction cost.

**Consumer price index (CPI)**

The consumer price index is a key economic figure and measures the change in prices of a collection/basket of goods for different periods and hereby inflation. House transactions are not included in this measurement although rent is a part of the basket. Rent is usually a factor relevant towards the house price and included in house price research, although there is an exception when it
comes to Denmark. The Danish rental market is strongly regulated, pro-tenant and some even describe it as crippled\(^3\). There are five different forms of rent control imposed by authorities\(^4\), hence, the rental market does not respond or reflect market forces (Skaarup & Bødker, 2010). For this reason rent will not be included as a variable. But it is a part of the CPI. In a stable house market, the prices should, in theory, adjust itself to the inflation or the CPI. Bork and Møller (2012) found it relevant to use in their model in forecasting US house price and used it in their three factor model to forecast the house price.

*Consumer confidence index*

The last independent variable is not a macro or microeconomic factor but a psychological behavior variable of the consumers. As mentioned in section 3, the significance of the consumer’s physiological behavior was brought up on the agenda during the recent housing bubble. Recently many researchers within the field of housing have highlighted and studied the financial behavior of people in the housing market. According to them, irrational and speculative behavior has had a significant and growing effect on the house price in the recent years (Shiller, 2008; Case & Shiller, 2004; Lunde, 2007; Mortensens & Seabrook, 2009). However the step from recognizing this as a relevant factor to actually measuring its effect on the house price is large. Finding tangible data on people’s expectation that can be placed and tested in a model is easier said than done. Dieci & Westerhoff (2012) developed a model on how a highly speculative market might affect the house price. The speculative forces in this case interfered with the real demand in the market lead to complex price dynamics that are capable of creating bubbles.

This thesis has been given access to data from a unique research on Danish people’s expectations towards the house price. In 2005, in the heat of the housing bubble as housing became more and more relevant and took more public space (reflected in tv-programs, newspapers etc), Green Analyse Institute, part of Børsen started a question survey. Including over 1000 participant above the age of 18 they asked the following question: “Do you expect that the housing prices will increase or decrease in the following 12 months?”, were the participants could answer “increase”, “decrease”, “unchanged” or “don’t know”. The survey was at first conducted irregularly on different occasions but in 2008, they started conducting this survey monthly. This data is unique as

\(^3\)www.globalpropertyguide.com, 2011
\(^4\)www.globalpropertyguide.com, 2011
there is no other institute or organization in Denmark that collects data on the expectations towards the house price on a regular basis. It provides the opportunity to statistically test this data against the house price and see if any significance can be found. The data has been named “consumer confidence index” as it describes the confidence consumers have in the housing market.

Average income family & Average personal income

The European Central Bank (2003) writes in their report that household income is the single most significant explanatory variable of housing prices. The reason why income changes have such a strong effect on the housing prices is partly human psychology, when people face a growing income, their demand for living and recreational space increases. This drives the prices on housing further up than the general price increases on goods that follow long run income increases (ECB, 2003). The long run elasticity of housing price with regards to income has according to empirical studies been close to unity (ECB, 2003). Skaarup & Bødker (2010) use income as an important variable in their model in the search for the price equilibrium on the Danish housing market. In their research paper on housing price forecasting Bork & Møller (2012) includes income in their three factor model that they are able to explain 50% of the variation in the house price, one-quarter ahead.

To surely capture the impact of this important variable this thesis will include two types of income variable in the model. The first variable includes the average income for all family types in Denmark or in other words the households. They are especially relevant since over 50% of all families are owner occupiers, they own houses (Lunde, 2009). The second variable includes data on the average private income for all people in Denmark. This is the average income for one person, and not a family or a household. Case & Shiller (2004) concluded that the income variable had a significant and large impact on the housing price in the majority of the states that they did their research on. A positive correlation between and a significant relationship are expected between these two variables and the housing price. Including two types of income as a variable will increase the chance of capturing a valid relevance between this important factor and the housing price.
5. Data collecting and sampling

5.1 Dependent variable

\[ Y = \text{House price index; houses and apartments} \]

The housing price index is adapted from Statistics Denmark, an official central organization on Danish statistics. This is a large, well-acknowledge, reliable and independent organization that follows the code of conduct for European statistics (Stat Denmark, 2013). The house price index that they produce is widely used within the housing industry, among banks and other organizations (Stat Denmark, 2013).

As stated in the delimitation, house price index for one-family will be used for the analysis. The frequencies of house price index vary between the years in Statistics Denmark’s databank. Yearly and quarterly data is available from 1992 to 2012 and from 2006 they have produced monthly house price indexes. Hence, quarterly observation was chosen because the quarterly observations 1992:Q1-2012:Q3 offers the most data compare to the other two, which is a total of 83 observations. Furthermore, this is important because analyze time series from two last decades will able to see a clearer behavior in the trend of the housing market in Denmark compare to analyzing monthly for only six years period.

5.2 Independent variables

Figure 5.2.1 Graphs collection of independent variables
From the top left and down: Construction cost, Consumer confidence index, Average personal income, Unemployment, Land price, Growth change.

From the top right and down: Consumer price index, Average family income, Population, Interest rate, Total Employment.

Graph collection 5.2.1 shows all the time series collected for the different independent variables. Since the dependent variable are collected from 1992 to 2012 quarterly therefore, for most of independent variables will be collected from 1992 to 2012. It will be stated otherwise if different period are collected.
5.3 Data Adjusted

To make it easier for analyze the time series regression, it is important to adjust the independent data so that all data have the same frequency with the dependent variable, which is quarterly.

As mentioned above, interest rate data is not quarterly but monthly therefore, to make it quarterly, taking average of every one quarter. Q1= January, February, March……Q4=October, November, December. The same will apply for the consumer confidence index for a period from 2008 to 2012.

For yearly data such as employment, population and consumer price index will be assumed that the whole year is the same for every quarter. In table 1 below, the variables are defined in the model of the thesis.

Table 5.3 The name of the variables when using in R

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>Unemployment</td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>Population</td>
<td></td>
</tr>
<tr>
<td>x3</td>
<td>Land price</td>
<td>The data from Statistics Denmark available until 2010:Q4.</td>
</tr>
<tr>
<td>x4</td>
<td>Interest rate</td>
<td></td>
</tr>
<tr>
<td>x5</td>
<td>Growth change</td>
<td></td>
</tr>
<tr>
<td>x6</td>
<td>Total employment</td>
<td>Full time, part time etc.</td>
</tr>
<tr>
<td>x7</td>
<td>Construction cost</td>
<td></td>
</tr>
<tr>
<td>x8</td>
<td>Consumer price index</td>
<td></td>
</tr>
<tr>
<td>x9</td>
<td>Consumer confidence index</td>
<td>The Green Analyse Institute started the survey in 2005 with irregular intervals, since 2008 regular surveys have been conducted monthly.</td>
</tr>
<tr>
<td>x10</td>
<td>Average family income</td>
<td>Available quarterly from 2000 to 2012, Statistics Denmark.</td>
</tr>
<tr>
<td>x11</td>
<td>Average personal income</td>
<td>Quarterly data from 2000 to 2012, Statistics Denmark.</td>
</tr>
</tbody>
</table>

Source: Authors
6. Hypothesis treatment

The main model of this thesis is using a combination of a linear regression model with an autoregressive integrated moving average model (ARIMA) model. The equation can be seen as follow:

\[ y_t = (\beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \ldots + \beta_k x_{k,t}) \\
+ (a_0 + a_1 y_{t-1} + \ldots + a_p y_{t-p} + \varepsilon_t + \delta_1 \varepsilon_{t-1} + \ldots + \delta_q \varepsilon_{t-q}) + d_1 + d_2 + e_t \]

Whereas,  
\[ y_t = \text{House price index in period } t \]
\[ t = \text{the current time period} \]
\[ d_1 = \text{Control variable for the time series period 2008:Q3} \]
\[ d_2 = \text{Control variable for the time series period 2009:Q1} \]
\[ e_t = \text{The error term of the model (the residuals)} \]

First bracket - Linear regression model:

- \( X_1 = \text{unemployment} \)
- \( X_6 = \text{Total Employment} \)
- \( X_2 = \text{Population} \)
- \( X_7 = \text{Construction cost} \)
- \( X_3 = \text{Land price} \)
- \( X_8 = \text{Consumer price index} \)
- \( X_4 = \text{Interest Rate} \)
- \( X_9 = \text{Consumer confident index} \)
- \( X_5 = \text{Growth change} \)
- \( X_{10} = \text{Average family income} \)
- \( X_{11} = \text{Average personal income} \)

\( x_{1,t} \ldots x_{k,t} = \text{Various independent variables from } x_1 \text{ to } x_{11} \text{ in period } t \)

\( \beta_0 \ldots \beta_k = \text{the parameters of independent variables to be estimated} \)
Second bracket - autoregressive integrated moving average (ARIMA):

\[ y_{t-p} = \text{the lagged value of dependent variables} \]

\[ a_0 \ldots a_p = \text{the parameter of the lagged value to be estimated} \]

\[ \varepsilon_{t-q} = \text{the error term of the dependent variables} \]

\[ \delta_1 \ldots \delta_q = \text{the parameters of the error term to be estimated} \]

The null hypothesis is that the parameters estimated in their model are not statistically different from zero. In order to test their significance, this thesis will consider an alpha level of p-value significantly at 0.05 (5%).
7. Methodology

The primary objective of time series analysis is to create a mathematical model that will provide an appropriate description of the sample data. In order to do this, it is assumed that a time series can be defined as a collection of random variables indexed according to the order they are obtained in time.

In Woodward (2012), the concept of stationary plays an important role in analyzing the time series. A process that is said to be stationary is when it is in a state of “statistical equilibrium” or the statistical properties remain constant over time. This means the basic behavior of a time series does not change over the course.

According to Woodward (2012), strictly stationary requires that for any \( t_1, t_2 \in T \) then the distribution of \( X(t_1), X(t_2) \) must be the same and furthermore, all bivariate distribution of pairs are the same for all \( h \). The requirement of strictly stationary is a severe requirement and usually is difficult to establish mathematically. Therefore, a common one of these were developed known as covariance stationary.

The time series said to be Covariance stationary if it follows four conditions:

1. Exhibits mean reversion in that it fluctuates around a constant long-run mean as 
   \[ E[X(t)] = \mu \text{ (constant for all t)} \]
2. Has a finite variance that is time-invariant as 
   \[ Var[X(t)] = \sigma^2 < \infty \text{ (A finite constant for all t)} \]
3. Has a theoretical correlogram that diminishes as lag length increases
4. \( \gamma(t_1, t_2) \) depends only on \( t_2 - t_1 \)

In this thesis, unless mention “strictly stationary”, stationary will be defined as “covariance stationary”. Thus, it will be assume that when mention stationary in this thesis, the series will satisfy the four conditions above.

7.1. Decomposing a time series

Before going into stationary of the time series, it is important to decompose a series into trend, seasonal, cyclical and an irregular component because uncovering the dynamic path of a series will improves forecast accuracy in Enders (1995). However, in many economic data, it is common to
find that a series contains a stochastic element in trend, seasonal and irregular components and since the working data is economic data so cyclical component will be unlikely to appear in the series.

Each of these components is a part of the series. The trend changes the mean of the series. The seasonal component exists when a series is influenced by seasonal factors such as quarter of the year, the monthly or daily so the seasonality is always a fixed or a known period. A cyclic pattern occurs when the data exhibit rises and falls that are not in a fixed period. As for irregular component, while it does not have any well-defined pattern, it varies between each time series but within that series, it is still predictable.

The possible equation as follow:

\[ y_t = T_t + S_t + C_t + I_t \]

Whereas, \( t \) stands a certain period of the time series

\( S_t \) stands for seasonal component in \( t \)
\( T_t \) stands for seasonal component in \( t \)
\( C_t \) stands for the cyclical component in \( t \)
\( I_t \) stands for irregular component in \( t \)

Each of these components has its own difference equation. A difference equation expresses the value of a variable as a function of its own lagged values, time and other variables. The trend and seasonal terms are both function of time and the irregular term is a function of its own lagged value with the pure random stochastic variable (Enders, 1995). Unlike the trend component, seasonal component happen with the a similar magnitude during the same period each year so it is not so interesting to analyze this equation such as ice cream seller, during the summer, the sales will increase greatly but during the winter, the ice cream man will not able to sell any. Possible pattern can occur in the housing therefore, if it is adjusted, it will not cause make the interpretation of the series to be ambiguous (Census, 2013). Therefore, seasonally adjusted can be done by taking the original time series and deducting the seasonal components out of the series.
7.2. Unit root test

There are big differences between stationary and non-stationary time series. A non-stationary series necessarily has permanent components. The mean and variance of a non-stationary time series are time-dependent. According to Enders (1995), the identification of a non-stationary series can be stated as follows:

1. There is no long-run mean to which the series returns
2. The variance is time-dependent and goes to infinity as time approaches infinity
3. Theoretical autocorrelations do not decay but in finite samples, the sample correlogram dies out slowly

This definition is not easy to spot out when using graphical techniques so in this thesis, an appropriate statistical test is conducted to check whether the series is stationary or not. And such, the procedure will be used is Dickey and Fuller unit root test (Enders, 1995).

Dickey and Fuller (1979) actually consider three different regression equations that can be used to test for the presence of a unit root:

\[
\Delta y_t = y y_{t-1} + \epsilon_t \quad (2)
\]

\[
\Delta y_t = a_0 + y y_{t-1} + \epsilon_t \quad (3)
\]

\[
\Delta y_t = a_0 + y y_{t-1} + a_2 t + \epsilon_t \quad (4)
\]

The difference between the three regressions concerns the presence of the deterministic elements \(a_0\) and \(a_2 t\). The first equation (2) is a pure random walk model whereas, the second equation (3) adds a drift term into the equation and the third equation (4) includes both a drift term and time trend.

The parameter of interest in all this regression equations is \(y\) as the hypothesis follow:

\(H_0: y = 0\) (The series has a unit root so the data needs to be differenced to make it stationary)

Comparing the observation value of \(t\)-statistic with the appropriate value reported in the Dickey-Fuller critical value tables to determine whether to accept or reject the null hypothesis of \(y = 0\).
It is highly possible to use the extensions of the Dickey-Fuller test, augmented Dickey-Fuller test. Not all the time-series processes can be well represented by the first-order autoregressive process so using this augmented Dickey-Fuller test will takes in account of the possible n-order. The statistical program, R when conducting a unit root test will use augmented Dickey-Fuller test instead of the normal Dickey-Fuller test because the software is taking account all of the possible order and if it is the first-order autoregressive process then the using augmented Dickey-Fuller will result exactly the same as the normal Dickey-Fuller test because in the augmented test, when n = 1 then the test becomes the normal Dickey-Fuller test.

It is appropriate to use Dickey-Fuller test statistic but it is important to know various issues in this test:

Before going into the issues of this test, when using the argument Dickey-Fuller test, it is assumed that that the series has a white noise and the appropriate lags will be put into the equation because if there is not enough lags then there will be serial correlation in the errors and if there are too much lags putting into the equation then the power of statistic goes down, to prevent this, BIC will be used to consider the best lags to use for the series.

The first issue is the power of a test is equal to the probability of rejecting a false null hypothesis. Monte Carlo simulations have shown that the power of the various Dickey-Fuller tests is very low. This means the unit root tests do not have power to distinguish between a unit root and a near unit root process. Thus, these tests will too often indicate the result that a series contains a unit root. Furthermore, it has little power to distinguish between trend stationary and drifting processes. By reducing the power means that it may conclude that the process contains a unit root when none is present (Enders, 1995).

According to Enders (1995), the second issue is the appropriate statistic for testing for H₀ because it depends on which regressors are included in the model. At a given significance level in the Dickey-Fuller critical table, the confidence interval around the H₀ expands if a drift and trend are included in the model. This is different when the series is stationary because the distribution of the t-statistic does not depend on the presence of the other regressors when stationary variable are used. The key problem is the tests for unit root are conditional on the presence of the deterministic regressors and testing for the presence of the deterministic regressors are conditional on the presence of a unit root.
7.3 Difference the equation

Making a time series stationary is important because many forecasting tools require the series to be stationary. According to (Duke Education, 2013), the stationary time series is easy to predict because its statistical properties will be the same in the future as it has been in the past.

It is possible to transform a non-stationary time series into stationary by using difference equations. Difference equations will normalize units so that \( h \) represents a unit change in \( t \) and consider the sequence of equally spaced value of the independent variables. It is possible to understand the first difference as the following equations:

\[
\Delta y_t = f(t) - f(t - 1) = y_t - y_{t-1}
\]
\[
\Delta y_{t+1} = f(t + 1) - f(t) = y_{t+1} - y_t
\]
\[
\Delta y_{t+2} = f(t + 2) - f(t + 1) = y_{t+2} - y_{t+1}
\]

This means, first difference of dependent values(y) is the difference of the present period value minus the previous period value and so on for all other values. Thus, it is also possible generate the 2\(^{nd}\) difference as follows:

\[
\Delta^2 y_t = \Delta(y_t - y_{t-1}) = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}
\]
\[
\Delta^2 y_{t+1} = \Delta(y_{t+1} - y_t) = (y_{t+1} - y_t) - (y_t - y_{t-1}) = y_{t+1} - 2y_t + y_{t-1}
\]

Therefore, \( n^{th} \) (whereas, \( n \) is the any positive integer such as 2, 3, 4, and 5 etc...) difference is defined analogously. The issue of whether the variables need to be stationary exists. According to Enders (1995), Sims (1980) and Doan (1992) recommend against differencing even if the variables contain a unit root. Because VAR determines the interrelationship among the variables, not the parameter estimates. The main argument against differencing is that it “throws away” information concerning the co-movements in the data (such as the possibility of co-integrating relationship). Similar, it is argued that the data need not to be detrended. In a VAR, a trending variable will be well approximated by a unit root plus drift. This is particularly true if the aim is to estimate a structural model. However, this thesis will take the approach of differencing the series to be stationary before conduct any further test. Although, there are possibility of losing information but
at the same time, forecasting the values can be more precise because forecasting a stationary time series will be easier to predict the patterns of the past year for the series.

After finishing this step, the data can be generated to be stationary. For an easier understanding, a graph can be made to compare the difference between the original data and the data after difference by n times.

7.4 Autoregressive integrated moving average model

Once the time series is stationary, there are possible time series model to understand the series such as GARCH, ARIMA. The focus of this thesis will be on Box-Jenkins (1976) methodology for estimating time-series model in the form of:

\[ y_t = a_0 + a_1 y_{t-1} + \ldots + a_p y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} \]

The equation above represents autoregressive integrated moving average, ARIMA (q,d, p) time-series models. Whereas, I(d) in ARIMA means the integrated by the d order.

The first part of the series \(a_0 + a_1 y_{t-1} + \ldots + a_p y_{t-p}\) is the autoregressive. The autoregressive model, AR (q) uses the past values of the dependent variables to explain the current value and q is the order of the model. This can be in classical sample as AR (2) means q is 2 so the model uses the past two values to explain the current value.

The second part of the series \(\varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q}\) is the moving average. The moving average, MA (p) uses the lagged values of the error (residuals) term to explain the current value of the dependent variable.

By using appropriate technique to find out the orders of this ARIMA model to create the main model of this thesis and forecast future values.

7.5 Combination of linear time series model

In the previous section, using ARIMA can capture many interesting dynamic relationships by using single equation time-series method. However, that alones might not be good enough to explain the model because the lagged values might not able to capture the entire dynamic in the independent variables as the variables themselves.
Therefore, the main focus of this thesis is to take a combination of the ARIMA model which founded previously and set up a linear regression model of the independent variables with its past value as well as other appropriate independent variables in one equation. This approaches will able to help the thesis to capture the possible dynamics of house price index as well as a clearer view on how the past literature finding on these independent variables to be relevant.

There are many approaches in conducting a multivariate time series such as intervention analyze, transfer function model or vector autoregressive (VAR). These are considered useful when using a combination of linear time series model. When it is not certain that a variable is actually exogenous, it is a good idea to let the time path of the dependent series \( (y_t) \) to be affected by the current and past value of the independent variables \( (z_t) \) and let the time path of the independent variables \( (z_t) \) to be affected by the current and past value of the \( (y_t) \) (Enders, 1995).

Intervention analysis can be used to see how many deterministic functions affect an economic time series. The intervention function is clear and simple to understand but there is a wide variety of possible intervention functions. The main assumption in intervention analysis is that the intervention function has only deterministic components. Transfer function analysis is appropriate if the intervention sequence is stochastic. However, with economic data it is not always clear that one variable is dependent and the others independent. Therefore, VAR (vector auto regression) treats all variables as jointly endogenous. Each variable is allowed to depend on its past values and the lagged values of all other variables in the systems.

However the approach of this thesis will make it a simple combination model so it will not exploit these theories.

There are few issues to consider when doing a combination model. There are two important difficulties involved in fitting a multivariate equation:

The first concern is the goal of fitting into a parsimonious model because a parsimonious model is preferable to a model with over-parameters. Enders (1995) mentioned that small samples encountered in economic data, a parsimonious model can be estimating an unrestricted model so severely which limit the degree of freedom as to render forecasts useless. Moreover, the possible inclusion of large but insignificant coefficients will add variability to the model’s forecast. However, in paring down the form of the model, two equally good models will likely arrive at two different transfer functions. As one model might have better fit in term of BIC or AIC then the other
model may have better residuals diagnostic properties. Furthermore, there is a potential cost to using a parsimonious model because as long as the independent variables are exogenous, the estimated coefficients and forecasts are unbiased even though the model is over parameterized such will not be the case if improperly imposes zero restriction on any of the polynomials in the model.

The second concern is the assumption of no feedback from the dependent variable to the independent variables. For the coefficient to be unbiased estimate of the impact effect of independent variables on the dependent variable then the dependent variables must be uncorrelated with the residuals (random error) at all lags (Enders, 1995). To understand further about the problem of feedback, an example can be given as follow: when consumer behaviors overact or underreact towards the change of the predict house price, it will create an observation of movement in the consumer behaviors much greater than the house price index. Therefore, it has to be assumed that the consumer behavior will be independent of themselves and not adjusted according to the prediction of the future dependent variables but changes by themselves, only then the actual model will be uncover.

To ease these issues, two models will be generated from this thesis. The first model will be keeping all the parameters for the forecasting and the second model will try to fit the parsimonious model. Note that when generate the model, all the independent variables will be difference until they become stationary. Furthermore, it is possible to find the significant of the independent variables for the model but their coefficient of these variables cannot be trust to tell the relationship because they are made to forecast not to tell the relationship between independent variables to dependent variables.

7.6 Model selection criteria

How well does the estimated model fit the data? Enders (1995) mentioned that adding additional lags in the ARIMA model (p,d,q) to the model will necessarily reduce the sum of squares of the estimated residuals but adding such lags entails the estimate of additional coefficients and an associated loss of degree of freedom. Moreover, including extraneous coefficients will reduce the forecasting performance of fitted model therefore; there are various model selection criteria that trade off a reduction in the sum of squares of the residuals for a more parsimonious model. The two commonly used model selection criteria are Akaike information criterion (AIC) and Bayesian information criterion (BIC) calculated as:
\[ AIC = T \ln(\text{residual sum of squares}) + 2 \cdot n \]

\[ BIC = T \ln(\text{residual sum of squares}) + n \ln(T) \]

Where \( n \) = number of parameters estimated; \( T \) = number of usable observation.

When creating lagged variables, some observations are lost so to make a fair comparison to other alternative models; \( T \) is always to be kept fixed (Ender, chapter 2).

The two models to aid for selecting the most appropriate model as the AIC or BIC should be as small as possible. When using the criteria to compare alternative models, each variable will increase the number of regressors then \( n \) also increases but should have an effect of reducing the residual sum of squares. Thus, if the regressor has no explanatory power, by adding it to the model will cause both the AIC and BIC to increase. Since \( \ln(T) \) will be greater than 2, the marginal cost of adding regressors is greater with the BIC than the AIC and since this thesis is working a lot with the linear model and ARIMA model so in this thesis, the BIC will be more appropriate to use.

Other than having two criterion (BIC and AIC), Box-Jenkins model selection is important to evaluate to select a good ARIMA model. Box and Jenkins popularized a three-stage method aimed at selecting an appropriate model for the purpose of estimating and forecasting a univariate time series.

In the identification stage, examines the time plot of the series, autocorrelation function and partial correlation functions. Plotting each observation of the dependent variable against time provides useful information concerning outliers, missing values and structural breaks in the data. Non-stationary variables may also have a pronounced trend or appear to meander without a constant long-run mean or variance. Missing values and outliers can be corrected at this point by difference any series deemed to be non-stationary.

In the estimation stage, a comparison of the sample ACF and PACF will provide several plausible models. This stage examines each of the models with various coefficients using a few criteria such as parsimony, stationarity and invertibility.

- Parsimony is defined as sparseness or stinginess. As mentioned earlier, incorporating additional coefficients will increase fit (the value of \( R^2 \) will increase) at a cost of reducing degree of freedom. Box and Jenkins argued that parsimonious models product better
forecasts than over parameterized model. The aim is to approximate the true data-generating process but not to pin down the exact process such as eliminating the MA (5) coefficient and takes a smaller model AR(2). In order to make sure the model is parsimonious, the coefficients should all have the t-statistic of 2.0 or greater. Moreover, the coefficients should not be strongly correlated with each other.

- Stationary as mentioned before, it is very important to have stationary before going into ARIMA
- The model is also should be invertible. Invertibility is important because the use of ACF and PACF implicitly assumes that the independent sequence can be well approximated by an autoregressive model.

The last stage is diagnostic checking. \( R^2 \) and the average of the residual sum of squares are common ‘goodness of fit’ measures in ordinary least squares. However, possible problem with these measures is that the fit is not necessarily improves as more parameters are included in the model so parsimony suggests using BIC as more appropriate measure of the overall fit of the model. The standard practice is to plot the residuals to look for outliers and evidence of periods in which the model does not fit the data well. The more modern tradition is to look for the white noise of the series. If the series has white noise that means the model is dynamically correct. White noise shall be discussed in the following section.

### 7.7 White noise test

In time series, uncorrelated data play an important role. The time series often refers to this as ‘White noise’. It also requires to satisfy three conditions:

1. The \( X_i \) are identical distributed
2. \( \gamma(t_1, t_2) = 0 \) When \( t_2 \neq t_2 \)
3. \( \gamma(t, t) = \sigma^2 \) Where \( 0 < \sigma^2 < \infty \)

Ljung and Box (1978) developed tests of a null hypothesis that the residual of the series are white noise. The Ljung-Box test statistic given as follow:

\[
L = n(n + 2)\sum_{k=1}^{K} \frac{\hat{\epsilon}_k^2}{n-k}
\]  
\[ (6) \]
The equation (6) is used to test the hypotheses:

$$H_0: \rho_1 = \rho_2 = \ldots = \rho_k = 0$$

Ljung and Box (1978) show that $L$ in the equation above (6) is approximately distributed as chi-squared with $K$ degrees of freedom when the data is white noise.

Alternative test for normality will be run by R such as Shapiro-Wilk test and Jarque-Bera test. These two tests will not be going into detail and will be discuss when use at result because these two are only to confirm whether the residuals follow normality. Once the model has white noise and the residuals follow normality, the model is dynamically correct equivalent to the model produces a reliable result.

7.8 Forecasting

There will be three models to forecast and compare in this thesis. First model will be the ARIMA found using the method in selection 7.4, the second model is the whole model including the ARIMA models with the independent variables and the last model is the model that would fit to be a parsimonious model. All these three will be forecast the same period for easier comparison.

After finding the combined regression models, the two models will be put into forecasting. This can be done by using the values of each independent variable and put it into the regression models to generate the dependent variable ($y$ value). The step is similar to a simple linear model. This means, without the value of the independent variables, it is not possible to forecast. Thus, it is not possible to forecast out of sample when using this techniques.

However, it is possible to forecast out of sample when using ARIMA model. Due to the limit of simplicity for this thesis, in depth of this out of sample forecasting will not be discuss because the main focus of this thesis is on finding out the dynamic behind the house price index. The forecast will take place but as one of the way to determine which model is more accurate and closer to the real world value.
8. Analysis and Result

8.1 ARIMA model:

When putting the data into R, the program can only read data from one to eighty three observations without any knowledge of what year were the data collected from or whether the data is in yearly, monthly or quarterly. It is important to convert the data into time series and this can be done by R statistical program and seen from the appendix A table 1. This table now shows that the data start from 1992 to 2012 in quarterly observation.

For a clear idea, a graph can be plot from this to observe the data:

Figure 8.1.1 House price index from 1992 to 2012

![House price index from 1992 to 2012](image)

Source: Authors (R-code generated in Appendix H Section 1)

In figure 8.1.1, the house price has increased greatly from 1992, but in the end of 2007, the financial crisis occurs and there has been a decrease in the house price ever since. As mentioned, the data should be stationary in order to develop a statistic model. However, looking at the figure 8.1.1, it seems that the data is likely to be non-stationary. This is only based on eye observation and cannot be determined with certainty so Dickey-Fuller unit root test will give a clearer knowledge on whether this data is stationary or not.

As mentioned in methodology, seasonality will be adjusted because keeping the seasonality will cause the interpretation of this series to be ambiguous. This can be done by to seasonally adjust the time series by subtracting the estimated seasonal component from the original time series. The estimated seasonal component can be generated by using R and seen from the appendix A figure 1.
There are two important points to take note of here:

First of all, there are some seasonal in the data so it is best to remove it. The closer look of the seasonal component graph can see as follow:

Figure 8.1.2: The graph of seasonal components in House price index variable

![Graph of seasonal components in House price index variable](image)

Source: by Authors

The graph in figure 8.1.2 is plotted against twelve months in 4 quarters. There can be seen some seasonality in this series. During quarter 2 and quarter 3, the price is at its peak which means the prices per house are higher during summer and autumn periods compare to quarter 1 and quarter 4.

Secondly, in random graph of appendix A figure 1, Quarter 3:2008 and Quarter 1:2009 are far off from other value so it will be putting as dummy variable to control for a better result. These two periods have shown a huge movement in term of error. As mentioned shortly in the methodology about the effect of feedback, these periods occurred after the financial crisis (the end of year 2007, the beginning of year 2008) it could interpret that Danish property market taken the after-shock effect. It is not clear when looking at the gathering data but due to the error components in the decomposition of house price; it was possible to foresee this big movement of the house price index.

In the appendix A figure 2, there is two graphs showing the data before and after the seasonal adjusted. At first glance, it is almost identical but when observes carefully, there are some small
differences between these two graphs as the seasonal variation has been removed from the graph. The series now just have the trend component and the random component.

After the data is seasonally adjusted, it is now possible to conduct a unit root test to check whether this time series data is stationary or not.

\[
\Delta Z_t = \beta_0 + \beta_1 t + \gamma Z_{t-1} + \sum_{i=1}^{\rho-1} \delta_i \Delta Z_{t-i} + e_t
\]

Whereas, $\beta_1$ is the deterministic trend and $\beta_0$ is the drift of the series.

The equation (8.1) shows the regression with auto-correlated errors considered for the unit root problem of a random walk with a drift term and deterministic trend. With this, the hypothesis of this test is

$H_0$: $\gamma = 0,$

The test result can be seen as follows:

Table 8.1.1 Test result of house price index for unit roots with six lags

| Coefficients: | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 1.19209  | 0.49436    | 2.411   | 0.01852 *|
| z.lag.1       | -0.03818 | 0.01752    | -2.179  | 0.03269 *|
| tt            | 0.03322  | 0.01862    | 1.784   | 0.07877 .|
| z.diff.lag1   | 0.99091  | 0.11064    | 8.956   | 3.19e-13 ***|
| z.diff.lag2   | -0.43959 | 0.14722    | -2.986  | 0.00389 **|
| z.diff.lag3   | 0.30975  | 0.11518    | 2.689   | 0.00894 **|

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple R-squared: 0.6646, Adjusted R-squared: 0.6406

Value of test-statistic is: -2.1791 2.1152 2.7174

Source: Authors (R command can be seen in appendix B result 1)
The table 8.1.1 is the unit root test is test for the trend of the time series and with BIC (Bayesian Information Criterion) to optimize the maximum lag of the time series. The BIC sets the lag to be at 3 instead of lag 6 so for a better test value, it is best to conduct the test once again with a lag 3 as follows:

Table 8.1.2 Test result of house price index for unit roots with optimize lags

<table>
<thead>
<tr>
<th>Test regression trend</th>
<th>Coefficients:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>1.26423</td>
</tr>
<tr>
<td>z.lag.1</td>
<td>-0.03902</td>
</tr>
<tr>
<td>tt</td>
<td>0.03320</td>
</tr>
<tr>
<td>z.diff.lag1</td>
<td>0.97549</td>
</tr>
<tr>
<td>z.diff.lag2</td>
<td>-0.43583</td>
</tr>
<tr>
<td>z.diff.lag3</td>
<td>0.31535</td>
</tr>
</tbody>
</table>

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Multiple R-squared: 0.6527, Adjusted R-squared: 0.6289
F-statistic: 27.43 on 5 and 73 DF, p-value: 1.625e-15

Value of test-statistic is: -2.2345 2.3594 2.9374

Source: Authors (R command can be seen in appendix B result 2)

After set the lag accordingly to the BIC, the table 8.1.2 shows the test statistic for the null hypothesis \( \gamma = 0 \) is -2.2345 and its corresponding critical value at level 1%, 5% and 10% are given in Appendix C table 1 as -4.04, -3.45 and -3.15 respectively. At these levels, it is impossible to reject the null hypothesis that \( \gamma = 0 \) and so it is possible to say that there is a unit root in this series. Furthermore, looking at the deterministic trend in the table 8.1.2, the test statistic of deterministic trend is 2.9374 with its corresponding critical value at level 1%, 5% and 10% are also given in Appendix C table 1 as 8.73, 6.49 and 5.47 respectively. This means the deterministic trend is not rejected and the trend term is not needed in this case. Since the trend is not needed so now it is important to consider the next equation for unit root test a random walk with drift.
\[
\Delta Z_t = \beta_0 + \gamma Z_{t-1} + \sum_{i=1}^{t-1} \delta_i \Delta Z_{t-i} + \epsilon_t
\]

Whereas, \(\beta_0\) is the drift of the series

The equation (8.2) shows the regression of a random walk with only a drift term. This hypothesis of this test is same as above. The procedure for R is also similar but instead of trend, it is now testing for the drift. Using BIC to also optimize the maximum lag of the time series as follows:

Table 8.1.3 Test result of house price index for unit roots with six lags

Test regression drift

\[\text{lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)}\]

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 0.731219  | 0.427922 | 1.709 0.0919 . |
| z.lag.1 | -0.008670  | 0.005865 -1.478 0.1438 |
| z.diff.lag1 | 0.981152  | 0.112194 8.745 6.95e-13 *** |
| z.diff.lag2 | -0.449767  | 0.149352 -3.011 0.0036 ** |
| z.diff.lag3 | 0.243218  | 0.110633 2.198 0.0312 * |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple R-squared: 0.6493, Adjusted R-squared: 0.6296
F-statistic: 32.86 on 4 and 71 DF, p-value: 1.682e-15

Value of test-statistic is: -1.4782 1.5344

Source: Authors (R command can be seen in appendix B result 3)

From table 8.1.3, the BIC optimizes the equation with lag 3, thus the model will rerun again with lag 3 instead of lag 6 and given the result as below:

Table 8.1.4 Test result of house price index for unit roots with optimize lags

Test regression drift

\[\text{lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)}\]

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|

44
(Intercept)  0.737708   0.398513   1.851   0.0681 .
z.lag.1   -0.008773   0.005586  -1.570   0.1206
z.diff.lag1  0.970700   0.111625   8.696 6.24e-13 ***
z.diff.lag2  -0.448848   0.149730  -2.998   0.0037 **
z.diff.lag3  0.252219   0.110859   2.275   0.0258 *

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple R-squared: 0.6368, Adjusted R-squared: 0.6172
F-statistic: 32.44 on 4 and 74 DF,  p-value: 1.305e-15

Value of test-statistic is: **1.5704 **1.8164**

Source: Authors (R command can be seen in appendix B result 4)

Looking at table 8.1.4. The test statistic for the null hypothesis $\gamma = 0$ is **1.5704** and its corresponding critical value at level 1%, 5% and 10% are given in Appendix C table 2 as -3.51, -2.89 and -2.58 respectively. Since it is not possible to reject the null hypothesis so from this test, it is possible to conclude that this series behaves like a random walk with a drift constant term.

From the test result, this time series data has a unit root so this means the data is non-stationary. Therefore, difference the data until it becomes stationary. From the test above, it shows that lag 3 is when the data has a unit root so difference the data by 3 will make it stationary.

**Figure 8.1.2** House price index graph (difference by 3)

![House price index stationary graph](image-url)
Looking at figure 8.1.2, the time series seems to be stationary as in mean and variance. Taking the first three differences in this time series, it has removed the trend component of the house price index and is left with a random component and so the time series should be stationary now. However, this is just eye observation of the graph so in order to make sure that the series is stationary, it is best to run the unit root test again to see whether after difference, the time series is now stationary or not.

In the appendix B result 5, the value of test-statistic is -11.1326 and its corresponding critical value 1%, 5% and 10% are given in appendix C table 1 as -4.04, -3.45 and -3.15 respectively. At this level, it is now possible to reject $H_0$; this means the series does not have a unit root and so the series is now stationary.

The next step would be selecting an appropriate model for this time series. As mentioned in the methodology, the model that will be using to analyze this time series is an ARIMA (p,d,q) model. Since the random components are the only left in this series because the series was differenced by 3 times in order to make it stationary, which means the trend component was removed when differenced. Furthermore, it can be assumed that I in (ARIMA) was 3 so the current stationary time series are without any difference. This means the model is left with ARMA and using the autocorrelation and partial autocorrelation to predict the p and the q in ARMA.

![Figure 8.1.3 The ACF graph of house price index differenced by 3](image)

Source: Authors (R-code generated in Appendix H section 1)
In figure 8.1.3 shows the graph of autocorrelation function (ACF) with its corresponding value, which can be found in appendix A, table 2. Looking at the figure 7.3, the autocorrelation at from lag 0 (1,000) to lag 2 (0.286) exceeds the significant boundary but all other between lag 2 to lag 5 do not exceed the significant bounds. Looking at partial autocorrelation function to find the order of the ARIMA model:

**Figure 8.1.4 The Partial ACF graph of house price index difference**

![Partial ACF graph of house price index difference](image)

Source: Authors (R-code generated in Appendix H section 1)

Figure 8.1.4 shows the graph of partial ACF for house prince index with its corresponding value, which can be found in appendix A, table 3. The partial autocorrelation shows that lag 4(-0.268) exceeds the significant boundary and it tails off to 0 after lag 4. However, there are some points taken here as before lag 1 as 0.25, 0.5 and 0.75 all three surpass the significant boundary and the same with before lag 2 so it is not exactly at lag 4 but this is slightly difficult to analyze the graph when only based on eye observation.

Since the correlogram shows at lag 2 and the partial autocorrelation shows at lag 4, so the following ARMA models are possible for this time series of third differences:

- An ARMA(4,0) model, so the AR(autoregressive) model is of order p = 4, since according to the partial autocorrelogram graph in figure 7.4, it is zero after lag 4 and the autocorrelogram graph in figure 7.3, when the lag is at 4, it is tails off to 0.
An ARMA(0, 2) model, so the MA(moving average) model is of order $q = 2$, according to the autocorrelogram graph in figure 7.3 because it is 0 after lag 2 and the partial autocorrelogram graph in figure 7.4, when the lag is at 2, it is tail off to 0.

An ARMA(p,q) is also possible in a situation where both of the model have the value greater than 0. However, since both of the graph in figure 7.3 and figure 7.4, both tail off to 0 so it is perhaps not an appropriate in this case.

With this, there are two models which can be a possibility as an ARIMA of the third differences but due to the fact that this graph is observe by looking so statistically speaking, it is not completely correct. Therefore, the best way in this case is using program R to conduct an auto ARIMA with set up of BIC (Bayesian Information criterion) to optimize for the best model.

Table 8.1.5 The result of Autoregressive integrated moving average (ARIMA) model

<table>
<thead>
<tr>
<th>Series: ydiff</th>
<th>ARIMA(3,0,0)(1,0,0)[4] with zero mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficients:</td>
<td>ar1      ar2      ar3     sar1</td>
</tr>
<tr>
<td></td>
<td>-0.7411  -0.7898  -0.7100  -0.4688</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.0831   0.0739   0.0864   0.1161</td>
</tr>
</tbody>
</table>

$\sigma^2$ estimated as 1.486: log likelihood=-130.49

AIC=270.98   AICc=271.79   BIC=282.89

Source: R-code generated in Appendix H Section 3

From this table 8.1.5, R gives the model an ARIMA model of (3,0,0) so this means Autoregressive model of third order. This model can be written as follow:

$$Y = \beta_1. Y_{t-1} + \beta_2. Y_{t-2} + \beta_3. Y_{t-3} + z_t$$

(8.3)

Whereas, $\beta_1, \beta_2, \beta_3$ are the parameters to be estimated and $z_t$ is random

So the house price index in Denmark is affected by the previous three quarters according to this model. From the test result, it is interesting to know that the current house price index is affected by the three previous quarters. And furthermore, the result produced above, all three quarters have a
negative correlation relationship, (this means when the house price index of the three quarters decreases then the current period of house price index increases).

From the equation 8.3, \( z_t \) could be a problem for the model because it is important to have the series to be white noise so that it does not have any effect on the other variables. By using R to test the random for the white noise of this ARIMA.

In appendix D result 1, the white noise test conducted for house price index autoregressive order of 3 gives Ljung-Box statistic = 12.93191, df = 25 p.value = 0.97732. This means, the random in this ARIMA model is white noise equivalent to the fact the model is dynamically correct.

Since using statistical test for ARIMA model, it shows that house price index in Denmark has a strong correlation with the three previous quarters. However, as stated in methodology, this paper wants to conduct a mixed regression model of the house price by combining linear regression model of independent variables with an ARIMA model.

8.2 Combination linear regression model

From the data collecting, there are eleven independent variables which have found to be relevant when analyzing the house price index. In the previous section 7, through test static, the house price index has an autoregressive order of 3 so a combination linear regression model can be form.

Before setting up a linear regression model, a few adjustments will be made for the independent variables so a possible good model can be generate. As similar to the analyses done above, Dickey Fuller unit root test will be used to find out whether these variables are stationary or not. For more accurate results, each variable will be seasonally adjusted. R commands for these steps can be seen in Appendix G section 2.

After finish with seasonal adjusted for each of the variables, next step is testing for each of these independent variables using argument Dickey Fuller unit root test, which can be shown in Appendix F, from result 1 to result 11. All of these tests have the same hypothesis for equation (8.1) and equation (8.2) above so there are 4 test results in each of them since both unit roots for trend and drift require testing separately with both lags adjusted afterwards. However, in result 5, there are only unit root tests for trend and the lag adjusted because the test statistic rejected the \( H_0 \), this means there is no unit root founded in independent variable (growth change), that equivalents to the time series of growth change is already stationary by itself no further necessary differencing.
Table 8.2.1: Requirement for Independent variables to be stationary

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Difference required for stationary</th>
<th>Variable name</th>
<th>Differences required for stationary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unemployment</td>
<td>4</td>
<td>Population</td>
<td>4</td>
</tr>
<tr>
<td>Land price</td>
<td>1</td>
<td>Interest rate</td>
<td>2</td>
</tr>
<tr>
<td>Growth change</td>
<td>Stationary</td>
<td>Total employment</td>
<td>1</td>
</tr>
<tr>
<td>Construction cost</td>
<td>1</td>
<td>Consumer price index</td>
<td>1</td>
</tr>
<tr>
<td>Consumer confident index</td>
<td>1</td>
<td>Average family income</td>
<td>1</td>
</tr>
<tr>
<td>Average personal income</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Authors

After conducting unit root tests for each independent variables (Appendix F result 1 to 11), it found the requirement of differences to make each of the independent variables stationary in the table 8.2.1 above. Unlike the method of VAR in methodology, these variables will be difference by the number required to be stationary instead of difference all these variables by one. Under Appendix G section 3, R scripts are shown on the program how to difference these variables.

It is now possible to create a combination linear model of the previous value of the house price index founded through ARIMA model (3,0,0) and different independent variables together in one model.

As mentioned it earlier, there are two dummy variables (d1 and d2) are included in the model to control for the two periods (Quarter 3:2008 and Quarter 1:2009) because using graphical technique, there is a huge jump between these two period of time so control for these will make the data more accurate. Below is the table to evaluate whether control variables should be use or not all for the model and the whole test result can be seen from Appendix E result 1 to result 4:
## Table 8.2.2  Model selection for the best controlled variables

<table>
<thead>
<tr>
<th>Nr</th>
<th>Dependent variable</th>
<th>The lagged values of dependent variables</th>
<th>Control variables</th>
<th>Independent variables included</th>
<th>Adjusted R-squared</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ydiff</td>
<td>$Y\text{diff}_t - 1 + Y\text{diff}_t - 2 + Y\text{diff}_t - 3$</td>
<td>No</td>
<td>$X\text{diff}_1, X\text{diff}_2, X\text{diff}_3, X\text{diff}_4, X\text{diff}_5, X\text{diff}_6, X\text{diff}_7, X\text{diff}<em>8, X\text{diff}<em>9, X\text{diff}</em>{10}, X\text{diff}</em>{11}$</td>
<td>0.8295</td>
<td>87.90202</td>
</tr>
<tr>
<td>2</td>
<td>Ydiff</td>
<td>$Y\text{diff}_t - 1 + Y\text{diff}_t - 2 + Y\text{diff}_t - 3$</td>
<td>$D_1, D_2$</td>
<td>$X\text{diff}_1, X\text{diff}_2, X\text{diff}_3, X\text{diff}_4, X\text{diff}_5, X\text{diff}_6, X\text{diff}_7, X\text{diff}<em>8, X\text{diff}<em>9, X\text{diff}</em>{10}, X\text{diff}</em>{11}$</td>
<td>0.8823</td>
<td>79.03892</td>
</tr>
<tr>
<td>3</td>
<td>Ydiff</td>
<td>$Y\text{diff}_t - 1 + Y\text{diff}_t - 2 + Y\text{diff}_t - 3$</td>
<td>$D_1$</td>
<td>$X\text{diff}_1, X\text{diff}_2, X\text{diff}_3, X\text{diff}_4, X\text{diff}_5, X\text{diff}_6, X\text{diff}_7, X\text{diff}<em>8, X\text{diff}<em>9, X\text{diff}</em>{10}, X\text{diff}</em>{11}$</td>
<td>0.8101</td>
<td>90.44244</td>
</tr>
<tr>
<td>4</td>
<td>ydiff</td>
<td>$Y\text{diff}_t - 1 + Y\text{diff}_t - 2 + Y\text{diff}_t - 3$</td>
<td>$D_2$</td>
<td>$X\text{diff}_1, X\text{diff}_2, X\text{diff}_3, X\text{diff}_4, X\text{diff}_5, X\text{diff}_6, X\text{diff}_7, X\text{diff}<em>8, X\text{diff}<em>9, X\text{diff}</em>{10}, X\text{diff}</em>{11}$</td>
<td>0.8842</td>
<td>79.07093</td>
</tr>
</tbody>
</table>

Source: Authors

From the table 8.2.2 above, the optimize model would be the one with the smallest value in BIC and highest value in adjusted R-square. There are two good models (2, 4) both underline in the table. The adjusted R-squared in model 4 is better than the model 2 but looking at the Bayesian information criterion (BIC), the value in model 2 is smaller than the model 4. Both of these two have equally close value but in this thesis, the model 2 (which is underlined and bold) is chosen to proceed further because the Bayesian information criterion is known for heavily penalize on over parameter equations but in this case, the Bayesian information criterion is smaller and also the adjusted R-squared of this model is almost the same to the adjusted R-squared of model 4.

There are two methods the thesis will look at. First, reduce some independent variables included in the model until the equation is optimized with a better adjusted R-squared and the BIC value. The second approach is to keep the whole model. Both of these approaches are valid because doing a multivariate time series model is very tricky as mentioned earlier in the methodology, one goal is to fit a parsimonious model but at the same time, the whole model with over parameters might possible give a better residuals diagnostic properties and forecast.
8.2.1 The model 1

From here on, this model will be referred as model 1. Below is the result of the whole model which will be used for forecasting in the later section:

Table 8.2.4 Result of the whole model (model 1):

| Coefficients | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------|----------|------------|---------|---------|
| Intercept    | -4.151e+00 | 2.218e+00 | -1.872  | 0.1104  |
| ydiff.1      | -4.303e-01 | 1.278e-01 | -3.366  | 0.0151 *|
| ydiff.2      | -1.463e-01 | 1.698e-01 | -0.862  | 0.4219  |
| ydiff.3      | -2.154e-01 | 1.655e-01 | -1.301  | 0.2409  |
| x1diff       | -4.996e-06 | 7.204e-06 | -0.693  | 0.5140  |
| x2diff       | -3.231e-05 | 2.710e-05 | -1.192  | 0.2782  |
| x3diff       | 1.170e-01  | 6.628e-02 | 1.765   | 0.1280  |
| x4diff       | -8.845e+01 | 1.183e+02 | -0.748  | 0.4828  |
| x5diff       | 2.395e-03  | 2.564e-01 | 0.009   | 0.9929  |
| x6diff       | 3.373e-03  | 9.668e-03 | 0.349   | 0.7391  |
| x7diff       | -5.303e-01 | 5.681e-01 | -0.933  | 0.3866  |
| x8diff       | 1.003e-01  | 5.552e-02 | 1.806   | 0.1209  |
| x9diff       | -6.716e+00 | 2.229e+00 | -3.013  | 0.0236 *|
| x10diff      | -1.874e-02 | 1.741e-02 | -1.077  | 0.3231  |
| x11diff      | 3.102e-02  | 2.869e-02 | 1.081   | 0.3211  |
| d1           | -2.522e+00 | 2.679e+00 | -0.941  | 0.3829  |
| d2           | 1.660e+01  | 7.216e+00 | 2.300   | 0.0611 .|

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Source: Authors (R-code can be seen in Appendix H section 3)

Looking at the table 8.2.4 above, it is founded that consumer confidence index(x9) is particularly significantly important to the model compare to other independent variables. From this, there seems to be an interesting link between the behaviors of consumer with the house price index.

With the model 1 being found, it is possible to do some residuals diagnostic properties to see whether the model has ‘white noise’.
Table 8.2.5 The normality test of model 1

<table>
<thead>
<tr>
<th>Shapiro-Wilk normality test</th>
</tr>
</thead>
<tbody>
<tr>
<td>data: resid</td>
</tr>
<tr>
<td>W = 0.9653, p-value = 0.5777</td>
</tr>
</tbody>
</table>

Source: Authors (R-code generated in Appendix H section 3)

This test is testing against the assumption of normality. Since the p-value is greater than the alpha level (0.05) causing to failure to reject the null hypothesis. However, fail to reject the null hypothesis does not mean it accepts that the data is normally distributed. This means it can only be assumed that the residuals are normality. Using graphical technical to help determine whether the model is normally distributed:

Figure 8.2.1 Q-Q plot of the model 1

Source: Authors (R-code generated in Appendix H section 1)

Looking at the graph 8.2.1, it does seem residuals follow the normality. Therefore, with the Shapiro-wilk test to fail to reject that the data is normality and with a Q-Q plot looks like the data is well-modeled by a normal distribution, it is expected that the data has ‘white noise’, which means the model is dynamically right.
8.2.2 The parsimonious model

The model that found in this section will be referred to as model 2. Looking at the result of the whole model above, there are a few variables which are seem to be insignificant in the model. For a good parsimonious model, the independent variable that is the most insignificant in the model will get removed slowly. After each variable are removed, the model will rerun again and check for the adjusted R-squared and the BIC to see whether they have improve. The controlled variables and the lagged value of dependent variables will not be change during the process of finding the parsimonious model. The table 2 below shows the result of removing the variables:

Table 8.2.6 Result for various models

<table>
<thead>
<tr>
<th>Nr</th>
<th>Dependent variable</th>
<th>The fix variables (no change to these)</th>
<th>Remove variables</th>
<th>Independent variables included</th>
<th>Adjusted R-squared</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ydiff</td>
<td>Ydiff$<em>{t-1}$ + Ydiff$</em>{t-2}$ + ydiff$_{t-3}$ + d1 + d2</td>
<td>None</td>
<td>X1diff, x2diff, x3diff, x4diff, x5diff, x6diff, x7diff, x8diff, x9diff, x10diff, x11diff</td>
<td>0.8823</td>
<td>79.03892</td>
</tr>
<tr>
<td>2</td>
<td>Ydiff</td>
<td>Ydiff$<em>{t-1}$ + Ydiff$</em>{t-2}$ + ydiff$_{t-3}$ + d1 + d2</td>
<td>X5diff</td>
<td>X1diff, x2diff, x3diff, x4diff, x6diff, x7diff, x8diff, x9diff, x10diff, x11diff</td>
<td>0.8991</td>
<td>75.90376</td>
</tr>
<tr>
<td>3</td>
<td>Ydiff</td>
<td>Ydiff$<em>{t-1}$ + Ydiff$</em>{t-2}$ + ydiff$_{t-3}$ + d1 + d2</td>
<td>X6diff, x5diff</td>
<td>X1diff, x2diff, x3diff, x4diff, x7diff, x8diff, x9diff, x10diff, x11diff</td>
<td>0.9099</td>
<td>73.23037</td>
</tr>
<tr>
<td>4</td>
<td>ydiff</td>
<td>Ydiff$<em>{t-1}$ + Ydiff$</em>{t-2}$ + ydiff$_{t-3}$ + d1 + d2</td>
<td>X1diff, x6diff, x5diff</td>
<td>x2diff, x3diff, x4diff, x7diff, x8diff, x9diff, x10diff, x11diff</td>
<td>0.9133</td>
<td>71.93374</td>
</tr>
<tr>
<td>5</td>
<td>ydiff</td>
<td>Ydiff$<em>{t-1}$ + Ydiff$</em>{t-2}$ + ydiff$_{t-3}$ + d1 + d2</td>
<td>X1diff, x6diff, x5diff, x10diff</td>
<td>x2diff, x3diff, x4diff, x7diff, x8diff, x9diff, x11diff</td>
<td>0.9128</td>
<td>71.3409</td>
</tr>
<tr>
<td>6</td>
<td>ydiff</td>
<td>Ydiff$<em>{t-1}$ + Ydiff$</em>{t-2}$ + ydiff$_{t-3}$ + d1 + d2</td>
<td>X1diff, x6diff, x5diff, x10diff</td>
<td>x2diff, x3diff, x4diff, x8diff, x9diff, x11diff</td>
<td>0.9126</td>
<td>70.45393</td>
</tr>
<tr>
<td>7</td>
<td>ydiff</td>
<td>Ydiff$<em>{t-1}$ + Ydiff$</em>{t-2}$ + ydiff$_{t-3}$ + d1 + d2</td>
<td>X1diff, x6diff, x5diff, x10diff, x7diff, x11diff</td>
<td>x2diff, x3diff, x4diff, x8diff, x9diff</td>
<td>0.9087</td>
<td>70.32668</td>
</tr>
</tbody>
</table>
In table 8.2.6, there were 8 different models were tested to fit best with a parsimonious model. From model 1 to model 8, each variable which has the highest value of insignificant to the model are removed slowly. The test stopped at model 8 because from there on, the adjusted R-squared is greatly reduced and the BIC has very high value compare to all the previous models. Both model 4 and model 7 are underlined in the table because they have the best value in either adjusted R-squared or BIC value. The actual results can be seen in Appendix E part 2 from result 2 to result 8.

However, model 7 in table 8.2.6 above is bold and underline because it is the model that will be chosen in this thesis. There are a few reasons for choosing this model. First, the difference between adjusted R-square is only 0.046(0.9133 – 0.9087) which equivalent to model 2 is only 4.6% better in term of presenting the model as this is relatively low difference. Second, the BIC in model 7 is the best compare to any other models in table 2. And last but not least, according to Enders (1995), the model that fit to be a parsimonious model will have all the variables to be significant. Thus, looking at the Appendix E result 4 compare to result 7, model 7 has only one insignificant independent variable in the model whereas, model 4 has six insignificant independent variables in the model.

Below is the result of model 7 (will now refer for the rest of the thesis as model 2), which fitted to be a ‘parsimonious model’ in this thesis:

<table>
<thead>
<tr>
<th>8</th>
<th>Ydiff</th>
<th>Ydiff_{t,1} + Ydiff_{t,2} + ydiff_{t,3} + d1 + d2</th>
<th>X1diff, x6diff, x5diff, x7diff, x11diff, x3diff</th>
<th>x2diff, x4diff, x8diff, x9diff</th>
<th>0.7834</th>
<th>107.078</th>
</tr>
</thead>
</table>

Source: Authors
From the table 8.2.7, this model is very different compared to the original model. In this model, there are many variables which are significant towards the model. This means by reducing the parameters, the model was able to recognize other variables to be significant, which would fit more with the past literature and finding.

When compare model 1 and model 2 found above, there are few things which are relevant towards this thesis:

Population, interest rate, consumer price index and consumer confident index are significant towards the house price index. It is important to observe that both the model 1 and the model 2 have the consumer confident index to be very significant. However, in this model 2, it is found that population, interest rate and consumer price index also have some impact on the equation.

Although, the significant of the variables for the models are undeniable to be found useful, the interpretation of the coefficient is not reliable in these models. This is because both of these models do not take the lagged values of independent variables in account. Due to the simplicity discuss in the methodology, these models when made did not consider the lagged values such as land price (x3) might have a delay effect on the house price of the next two periods then the current value will not be significant important until the next three period. This could also be one of the reasons why
all the independent variables found in the past literature and findings show to be not significant in these two models.

The model 2 which fits to be a parsimonious model will also check for residuals diagnostic test.

Table 8.2.8 The normality test of model 2

<table>
<thead>
<tr>
<th>Shapiro-Wilk normality test</th>
</tr>
</thead>
<tbody>
<tr>
<td>data: resid</td>
</tr>
<tr>
<td>W = 0.9313, p-value = 0.117</td>
</tr>
</tbody>
</table>

Source: Authors (R-code generated in Appendix H section 3)

Figure 8.2.2 The Q-Q plot of the model 2

![Q-Q plot](image)

Source: Authors (R-code generated in Appendix H section 1)

The same test as above since the p-value is higher than the alpha level (0.05) so it is fail to reject the null(that the data is normality) but again this does not mean it accepts the data is normally distributed. However, with a Q-Q plot that looks like the residuals follow normality so it can only be assume to that the data is well modeled by normal distribution. With this, it can assume that the model has ‘white noise’.

Based on the generated results, it can be seen that model 2 seems to be the most feasible model compare to model 1 and the pure ARIMA model because it has the highest adjusted R-squared and
the lowest BIC value. The pure ARIMA model seems to be the worst model of the three models in comparison because it has a very high BIC in table 8.1.5. However, these are the first glance of the model based on model criterion. Thus, in the next section, this thesis will turn the model from statistical analysis to solving real world problem by making these models comparing with each other on how accurate the model can forecast.

8.3 Forecasting

In this section, using the three models founded in the previous sections to make a prediction of the house price index. Since it is a linear model, to be able to forecast for the future value of house price index, it requires the future values of the independent variables. Therefore, instead of predicting the future value, the model will instead predict the values that the data is available on and compare it to the real world data (the original data collected on house price index) to see how accurate it is. This way, it will help to determine which model is better when it comes to solving real world problem.

In the appendix F, result 1 contains the forecast result of the whole model; result 2 contains the forecast result of the parsimonious model and result 3 contains the forecast result of the ARIMA model. The forecast period was 2005:Q1 to 2010:Q3. Below is the table to compare to between these forecast values and the original value.

The original value that will be compared to is the house price index after seasonally adjusted and difference by 3. It will be underline in the table. The forecast values of these models will be taken within the confident of 95%. Whenever, the forecast values do not match with the original value in the table, those values will be in a small circle and bold.

<table>
<thead>
<tr>
<th>Period</th>
<th>The whole model</th>
<th>The original value</th>
<th>The parsimonious model (1,0,0)</th>
<th>ARIMA model(3,0,0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>2005:Q1</td>
<td>-2.9015</td>
<td>1.8251</td>
<td>0.6636</td>
<td></td>
</tr>
<tr>
<td>2005:Q2</td>
<td>-2.1744</td>
<td>2.3659</td>
<td>-0.2011</td>
<td>-1.3428</td>
</tr>
<tr>
<td>2005:Q3</td>
<td>-2.5498</td>
<td>2.2989</td>
<td>0.4360</td>
<td>-1.9075</td>
</tr>
<tr>
<td>2005:Q4</td>
<td>-3.1274</td>
<td>2.1935</td>
<td>-0.2985</td>
<td>-2.2257</td>
</tr>
<tr>
<td>Year:Q</td>
<td>Value1</td>
<td>Value2</td>
<td>Value3</td>
<td>Value4</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>2006:Q1</td>
<td>-2.8342</td>
<td>2.1969</td>
<td>-0.6363</td>
<td>-2.2534</td>
</tr>
<tr>
<td>2006:Q2</td>
<td>-4.4649</td>
<td>0.6219</td>
<td>-0.5011</td>
<td>-3.7972</td>
</tr>
<tr>
<td>2006:Q4</td>
<td>0.5131</td>
<td>5.8286</td>
<td>0.7014</td>
<td>0.8329</td>
</tr>
<tr>
<td>2007:Q1</td>
<td>-4.4287</td>
<td>0.3698</td>
<td>2.9636</td>
<td>-3.0198</td>
</tr>
<tr>
<td>2007:Q2</td>
<td>-1.3556</td>
<td>3.3409</td>
<td>-1.9011</td>
<td>-0.63137</td>
</tr>
<tr>
<td>2007:Q3</td>
<td>-2.7274</td>
<td>1.9185</td>
<td>0.6360</td>
<td>-1.7968</td>
</tr>
<tr>
<td>2007:Q4</td>
<td>-0.8196</td>
<td>4.4660</td>
<td>-1.1985</td>
<td>-0.03929</td>
</tr>
<tr>
<td>2008:Q1</td>
<td>-1.5321</td>
<td>3.7588</td>
<td>1.5636</td>
<td>-0.78660</td>
</tr>
<tr>
<td>2008:Q2</td>
<td>-6.6434</td>
<td>-1.2844</td>
<td>0.9988</td>
<td>-6.06532</td>
</tr>
<tr>
<td>2008:Q3</td>
<td>-4.1307</td>
<td>1.0568</td>
<td>-3.9639</td>
<td>-3.66054</td>
</tr>
<tr>
<td>2008:Q4</td>
<td>3.3841</td>
<td>8.7431</td>
<td>-1.4985</td>
<td>3.96230</td>
</tr>
<tr>
<td>2009:Q1</td>
<td>0.83261</td>
<td>5.9365</td>
<td>6.0636</td>
<td>1.5149</td>
</tr>
<tr>
<td>2009:Q2</td>
<td>-6.1645</td>
<td>-0.9340</td>
<td>3.2988</td>
<td>-5.4751</td>
</tr>
<tr>
<td>2009:Q3</td>
<td>-3.3352</td>
<td>1.5857</td>
<td>-3.8639</td>
<td>-2.6928</td>
</tr>
<tr>
<td>2009:Q4</td>
<td>-2.7133</td>
<td>2.1689</td>
<td>-1.3985</td>
<td>-2.7577</td>
</tr>
<tr>
<td>2010:Q1</td>
<td>-1.9291</td>
<td>2.4657</td>
<td>0.3636</td>
<td>-1.5956</td>
</tr>
<tr>
<td>2010:Q2</td>
<td>-4.4767</td>
<td>0.3177</td>
<td>0.6988</td>
<td>-3.5757</td>
</tr>
<tr>
<td>2010:Q3</td>
<td>-2.3681</td>
<td>2.4438</td>
<td>-1.8639</td>
<td>-1.1925</td>
</tr>
</tbody>
</table>
Looking at the table above, it seems that the forecast values are not accurate during 2008:Q2 to 2009:Q3. It could be because of the financial crisis occurred in the end of 2007 and the beginning of 2008. In the two models (1, 2), there were no actual variables which could account for the duration or the after-shock effect of the financial crisis. There are two dummy variables which controlled for two specific periods but the financial crisis does not only occur in these two periods as there are the pre-shock as well as the after-shock effect from the financial crisis or the bubble burst.

The model 1 is better at forecasting compare to the parsimonious model as it has 58.3 % correct during the 24 periods while model 2 has only 45.8 % correct in forecasting these period and the ARIMA model is the best model to forecast during these periods, as having 79.1 % correct out of the 24 periods forecasted.

Since the ARIMA model seem quite accurate in the previous table, it is possible to predict the future value of house price index such as in 2013:Q4 to 2015:Q4.

Figure 8.3.1 Forecast of using ARIMA model for the period 2012:Q3 to 2014:Q4

In the figure 8.3.1, the circle area shows the forecast period using the ARIMA model. The actual values of forecasting can generate through the graph in Appendix F result 4. It is important to investigate
whether the forecast errors of this ARIMA model are normally distributed with mean zero and constant variance and whether there are correlations between the forecast errors because this will help to show how accurate these values are:

Figure 8.3.2 The ACF graph of the forecast value

Source: Authors (R code can be generated in Appendix H section 1)

Table 8.3.2 the white noise test for ARIMA model

Test for independence of residuals

<table>
<thead>
<tr>
<th>Ljung-Box first order independence test</th>
</tr>
</thead>
<tbody>
<tr>
<td>data: resid</td>
</tr>
<tr>
<td>Ljung-Box statistic = 12.92957, df = 25</td>
</tr>
<tr>
<td>Null hypothesis: rho(1) = rho(2) = ... = rho(25) = 0</td>
</tr>
</tbody>
</table>

Source: Authors (R-code generated in Appendix H Section 3)

Since the correlogram shows that none of the forecast values autocorrelation from lag 1 to lag 5 exceeds the significant bonds and the p-value for Ljung-Box test is 0.977, it is possible to conclude that there is very little evidence for non-zero autocorrelation in the residuals of the forecast values. To see whether the residuals are normally distributed with mean zero and constant variance, plotting the time series graph and histogram will give a clearer idea:
From figure 8.3.3, the time plot of the residuals in the forecast shows a variance which seems to be roughly constant over time and the histogram in figure 8.3.4 seems to be normally distributed as it seemed to confirmed by Ljung-Box test.

Since the residuals of the forecast values do not seem to be correlated and normally distributed so the ARIMA model(3,0,0) with seasonal(1,0,0) does seem to provide an adequate predictive model of house price index.
Overall, the model 2 satisfy statistical properties best in term of adjusted R-square and the Baysian information criterion (BIC) but when it comes to solving real world problem, the pure ARIMA shows a more accurate result of forecasting compare to the others two models. Looking closely at the range of value predicting, the model 2 has the smallest range of values therefore, it could be explain why it has the lowest accuracy in forecasting. Although model has over-parameters but because of that, the range of values are bigger, thus, the accuraccy of forecasting is slightly better than model 2. The ARIMA forecasts the best out of the three models and this could be explain that in the ARIMA model, by having only the three lagged value of the house price index as the independent variables so the co-efficient of these values are higher and these lagged values have include and take account for all other variables, even the unexpected events such as shocks or financial crisis of the economy. Whereas, the two combinations models did not have these variables to take in account, there are two dummy variables which have taken in account some of these events but the financial crisis does not occur only in two periods and it could possible occur in pre-shock effect and the after-shock effect.
9. Conclusion

The purpose of this thesis was to create a model to explain the movement in the Danish housing price. Eleven independent variables were selected as the main factors on the Danish house price index, after a thorough research of the past literature and studies. These variables are economic fundamentals and demographic factors as employment, population, land price, growth change, total employment, construction cost, average personal and family income and consumer price index. There was also one new and untested variable that was included, i.e the consumer confidence index.

These variables were put up into a combination model between linear regression model and autoregressive integrated moving average (ARIMA) model. Thus, the analysis was split into two parts: the finding of ARIMA model and the finding of combination regression model.

While finding the appropriate model for ARIMA model, two dummy variables to control for the specified periods (2008:Q3, 2009:Q1) were created. These two represent a huge movement in the error which could be explained by the possible after-shock effect from the financial crisis. The time series (house price index), after having conducted augmented Dickey-Fuller unit root test, was found to be non-stationary. Therefore, the series was differenced by 3 to make it stationary.

The ARIMA model was found to have an autoregressive (AR) order of 3. Statistically, part of the Danish house price index can be explained by the previous three quarters.

The first part of the testing found the ARIMA model to have an autoregressive of 3. In order to find a combination regression model, each independent variable was differenced to a stationary form before putting it in the combination regression model.

After having evaluated through the Bayesian information criterion and the adjusted R-square, two possible models were found. The first model (model 1) is the whole model which includes the lagged values of the house price index in the last three quarters with eleven independent variables and two dummy variables. In this model, it shows that all the independent variables, except for consumer confidence index, are insignificant in this series. With the model having adjusted R-squared to be 0.8823, this means that 88.23% of the variations in the house price index could be explained by the independent variables in the model (eleven factors and the previous three quarters of house price index).
The second model (model 2) was made to test if a parsimonious model would resolve in a stronger model. By removing the non-significant independent variables one by one, the final model came to include the lagged values of house price index with five independent variables and two dummy variables. In this case, it shows that *interest rate, population, consumer price index and consumer confident index* are significant to the model. In this model, the adjusted R-squared is 0.9087 shows that 90.87% of variation in the house price index can be explained by the independent variables in this model (five factors and the previous three quarter of house price index).

Both model 1 and model 2 found the ‘consumer confident index’ variable to be significant. This is a unique finding since this is the first time this variable gets tested towards the house price index. It has the most significant impact in describing the house price index among of the eleven independent variables. In other words people’s expectation towards the hose price has a relevant effect on the house price movements.

Although, the model 2 has the best statistical properties (high adjust R-square, lowest BIC value), the model 1 was still chosen as one of the best models because according to Enders (1995), having restriction on the model will limit the ability to forecast. The two models and the ARIMA model’s fit to forecast were tested through an in-sample forecasting during the period from 2005:Q1 to 2010:Q3. Model 1 was more accurate than the model 2 in range of forecasting. It has 58.3% correct during the 24 periods that it forecasts whereas, the model 2 only has 45.8%, which means more than half of what it predicts is incorrect. The ARIMA model is the best model to forecast during these periods, as having 79.1% correct out of the 24 periods it forecasts.

All three models use the lagged value of the house price index but model 1 and model 2 also include the independent variables and yet they predict worse than the ARIMA model (without any independent variables). Does this mean the model with the least parameters predict better? It is not the case either because model 1 has better accuracy in forecasting than model 2. Furthermore, in order to simplify the problems; this thesis did not include the lagged values of each of the independent variables.

It can be concluded that the ARIMA model is the most accurate when it comes to forecasting. An attempt to forecast the house price for the next 9 periods was made (from 2012:Q4 to 2014:Q4). It turns out that the house price index in 2012:Q4 newly available from StatBank (Statistics Denmark)
difference by three gives a result of 0.8218. This value lies within the interval of the ARIMA model forecasted (from -2.5136 to 2.2652).

9.2 Further improvement

One of the problems which could have improved better for this thesis is the consideration of the lagged values of independent variables. Using the VAR model, it could help in finding the appropriate lagged values of the independent variables on the house price index. This could also be the reasons to why the independent variables that were in the past literature found significant important to their models but were not significant in this model such as income might have a possible delay effect because a rise in family income will have higher chance of purchasing the house in the next period.

Furthermore, one could discuss that the house price index has been differencing it by three times is too many. As the more the difference, the more possible information might be thrown away. This applies for all the independent variables as well. Applying VAR model, will only difference all the variables by 1 but using more advanced and sophisticated methods contradicting to the goal of this thesis to have an open and simple approach to the housing price.

Different kinds of mortgages might play an important role in describing the housing price too. However, these factors are very difficult to collect data.
References

Articles and books are to be referenced as follows:


Det økonomiske råd (1979): “DANSK ØKONOMI – Dansk økonomi i 1970-80 og de nærmest følgende år” nr.36


European Central Bank (2003): ”Structural Factors in the EU Housing Markets”


Haagerup, C. (2009): “Tobin’s q for Danish single-family houses” Danish Building Research Institute, University of Aalborg


Lund, J. (2008): ”The housing price downturn and its effects – a discussion” Working paper 2008/1 Department of finance, Copenhagen Business school


Mattson, H.


Mortensen, J. L. Seabrooke, L. (2009)”Egalitarian Politics in Property Booms and Burst:: Housing As Social Right or Means to Wealth in Australia and Denmark”


Realkredit Danmark (2012): ”Året der gik på boligmarkedet” December 11.


Sutton, Gregory (2002): ”Explaining changes in house price” Bis quarterly review


Økonomi- og Erhvervsministeriet (2008): “Afdragsfrie lån og prisstigninger på boligmarkedet” NOTAT, 16 september

Internet sources are to be referenced as follows:

AFP article, published in 2007, Housing woes take bigger toll on economy than expected: Paulson, Washington, Copyright 2013 AFP, viewed 30 April 2013, http://afp.google.com/article/ALeqM5hWSjWmGJ4YXTh3Pm5kOC7csTT48g


Census Government, FAQs on seasonal adjustment, Viewed 30 April 2013, http://www.census.gov/const/www/faq2.html#five

Den Store Danske, Kartoffelkuren (The potato reform), Verify article, Viewed 30 April 2013, http://www.denstoredanske.dk/Samfund_jura_og_politik/%C3%98konomi/Pengeinstitutter_mv./Kartoffelkuren?highlight=kartoffelkur


Global Property Guide, Denmark’s short-lived housing market recovery, Published in December 2011, viewed 30 April 2013, http://www.globalpropertyguide.com/Europe/Denmark/Price-History
Total kredit, Hvor meget kan du lane?, Viewed 30 April 2013,
http://www.totalkredit.dk/page?action=page&pageId=267

Value Walk, Denmark and USA housing Bubbles have similarities, Published April 2012, Viewed 30 April 2013,
http://www.valuewalk.com/2012/04/denmark-and-usa-housing-bubbles-have-similarities/