CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Ondřej Šváb

Best predictors of apartment prices: Empirical Evidence from Czechia

Bachelor thesis

Prague 2019

Author: Ondřej Šváb **Supervisor**: Petr Pleticha M.Sc.

Academic Year: 2018/2019

Bibliographic note

ŠVÁB, Ondřej (2019). *Best Predictors of Apartment Prices: Empirical Evidence from Czechia*, Prague. 51 pp. Bachelor thesis (Bc.) Charles University, Faculty of Social Sciences, Institute of Economic Studies. Thesis supervisor Petr Pleticha M.Sc.

Abstract

It is essential to control for property price determinants since there could be created the price bubble, and its burst would have harmful effects on the economy. Thus, this bachelor thesis aims to show the best determinants and models for forecasting the apartment prices in Czechia and its regions with the use of panel data and time series from the Czech Statistical Office. After stating hypotheses of variable's expected impacts on apartment prices, the most important determinants appeared to be the average wage, unemployment rate, natural population growth, and the building plot price. The best results are found by using econometric regressions as the fixed effects, the first differences or the classical ordinary least squares method. I also use the heteroskedasticity and autocorrelation consistent standard errors for better robustness of coefficients. Moreover, the lasso method is applied for dealing with multicollinearity and over-fitting, which are fixed by the variable selection. In most cases, the lasso improved prediction accuracy. However, the first difference regressions worsen the forecasts after the lasso penalisation.

Keywords

apartment price, Czech housing market, price determinants, prediction, time series, panel data, lasso

Abstrakt

Kontrolovat ukazatele cen nemovitostí je klíčové, jelikož může docházet k tvorbě cenové bubliny, jejíž prasknutí by mohlo mít velice negativní dopady na hospodářství. Cílem této bakalářské práce je najít nejlepší možné ukazatele a modely pro předpovídání cen bytů v České republice a jejích regionech s užitím panelových dat a časových řad poskytnutých Českým statistickým úřadem. Po stanovení hypotéz o očekávaném vlivu proměnných na cenu bytů se průměrná mzda, míra nezaměstnanosti, přirozený růst populace a cena pozemku zdají být nejdůležitějšími ukazateli cen bytů napříč všemi modely. Optimální výsledek je dosáhnut pomocí ekonometrických modelů jako jsou fixní efekty, první diference nebo klasická metoda nejmenších čtverců s heteroskedasticitními a auto-korelovanými konsistentními standartními chybami. Dále je použita lasso metoda pro řešení multi-kolinearity a selekci velkého počtu nezávislých proměnných. Lasso metoda vylepšila přesnost předpovědi ve většině případů. Avšak prognóza regrese první diference se po aplikaci lasso penalizace zhoršila.

Klíčová slova

cena bytů, český trh s nemovitostmi, cenové ukazatele, prognóza, časová řada, panelová data, lasso

Declaration of Authorship

I hereby proclaim that I wrote my bachelor thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used.

I grant a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, 30 July 2019

Signature

Acknowledgment

I would like to express my gratitude to my supervisor, Petr Pleticha MSc., who provided me with valuable insights and helped me to shape this thesis. I would also like to thank Bc. Jan Malecha for his helpful comments. Last, but not least, I am thankful to my family and friends for their support during the studies.

Contents

List of Tables and Figures **i**

Acronyms **ii**

Thesis Proposal **iv**

1 Introduction **1**

2 Literature Review **4**

3 Data Description **9**

3.1 Data Structure 9

3.2 Hypothesis Statement 11

3.3 Apartment Prices 11

3.4 Supply Side Variables 12

3.5 Demand Side Variables 13

3.6 Summary of Descriptive Statistics 16

4 Methodology and Empirical Models **17**

4.1 Panel data approach 17

4.2 Time series approach 20

4.3 Penalised Least Squares approach 24

5 Empirical Results **28**

5.1 Panel Data Approach 28

5.2 Time Series Approach 32

5.3 Best Predictors of apartment prices 37

6 Conclusion **39**

Bibliography **41**

Appendix **44**

List of Figures

1.1 Real prices of second-hand apartment prices (100=2000) 3
6.1 Deflection of Lasso prediction for Time-Demeaning Data from the Actual Apartment Price 44
6.2 Deflection of First Difference Prediction for Time Series from the Actual Apartment Price 44
6.3 Cross-Validation for the LASSO Regression (FD Panel Data) 45
6.4 Number of Non-Zero Coefficients Depending on the Value of lambda (FD PD) . 45
6.5 Cross-Validation for the LASSO Regression (FE) 46
6.6 Number of Non-Zero Coefficients Depending on the Value of lambda (FE) 46
6.7 Cross-Validation for the LASSO Regression(OLS) 47
6.8 Number of Non-Zero Coefficients Depending on the Value of lambda (OLS) . . . 47
6.9 Cross-Validation for the LASSO Regression(Detrending) 48
6.10 Number of Non-Zero Coefficients Depending on the Value of lambda (Detrending) 48
6.11 Cross-Validation for the LASSO Regression (Seasonality) 49
6.12 Number of Non-Zero Coefficients Depending on the Value of lambda (Seasonality) 49
6.13 Cross-Validation for the LASSO Regression (FD Time Series) 50
6.14 Number of Non-Zero Coefficients Depending on the Value of lambda (FD TS) . . 50
6.15 Housing Price Index Growth and Interest Rate Changes 51
6.16 Housing Price Index and GDP per Capita Growth (100=2000) 51

List of Tables

3.1 Explanation of Variables 10
3.2 Descriptive Statistics - Panel Data 16
3.3 Descriptive Statistics - Time Series Data 16
4.1 Hausman test for panel data 18
4.2 Breusch-Godfrey test for panel data 19
4.3 Unit Root Test for Stationarity 21
4.4 Serial Correlation of Time Series 22
4.5 The Variance Inflation Factor for Panel data 25
4.6 The Variance Inflation Factor for Time Series 25
5.1 Panel Data and Lasso Regressions: Summary Tables 30
5.2 Time Series Regressions: Summary Tables 33
5.3 Lasso Regressions for Time Series Models 34
5.4 Root Mean-Squared Errors for Regressions 37

Acronyms

AR	Autoregressive Model
CEE	Central and Eastern Europe
CNB	Czech National Bank
CZK	Czech Crown
CZSO	Czech Statistical Office
FD	Fixed Effects
FDI	Foreign Direct Investments
GDP	Gross Domestic Product
GLS	Generalised Least Squares
HAC	Heteroskedasticity and autocorrelation consistent (standard errors)
IRI	Institute of Regional Information
LASSO	Least Absolute Shrinkage and Selection Operators
MFCR	Ministry of Finance of the Czech Republic
MRDCR	Ministry of Regional Development of the Czech Republic
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
PLS	Penalised Least Squares
RE	Random Effects
RMSE	Root Mean-Squared Error
SE	Standard Error
TS	Time Series
VIF	Variance Inflation Factor

Acronyms of Variables and Regions

apartp	Apartment Price
avgwage	Average Wage
bplotp	Building Plot Price
divorce	Divorce Rate
FDI	Foreign Direct Investments
finapart	Finished Apartments
gdpcap	GDP Capita
intrat	Interest rate
marriage	Marriage Rate
migr	Net Migration
popdens	Population Density
popgrwth	Population Growth
strapart	Started Apartments
unemp	Unemployment Rate
1	Praha (Prague)
2	Středočeský kraj (Stredocesky Region)
3	Jihočeský kraj (Jihocesky Region)
4	Plzeňský kraj (Plzensky Region)
5	Karlovarský kraj (Karlovarsky Region)
6	Ústecký kraj (Ustecky Region)
7	Liberecký kraj (Liberecky Region)
8	Královéhradecký kraj (Kralovehradecky Region)
9	Pardubucký kraj (Pardubicky Region)
10	Vysočina (Vysocina)
11	Jihomoravský kraj (Jihomoravsky Region)
12	Zlínský kraj (Zlinsky Region)
13	Omoloucký kraj (Olomoucky Region)
14	Moravskoslezský kraj (Moravskoslezsky Region)

Thesis Proposal

Author:	Ondřej Šváb
Supervisor:	Petr Pleticha, MSc.
Proposed topic:	Bubbling Real Estate Markets:
	An Empirical Analysis from the Czech Republic

Research Question and Motivation

I would like to study the growing prices of a real estate that has been rapidly growing last few years in many European metropolies as well as in the Czech Republic. More precisely said, if we talk about the Czech Republic, we focus on Prague and other major towns of particular districts in the country. The growth of prices will be compared with other macroeconomic determinants as unemployment or economic growth. The significance of this work is given by the possibility of the real estate market's overheating. It could lead to the creation of price bubbles which are harmful for the economic stability. The price bubble in real estate markets contributed to the Great recession after the steep decline in housing prices. Since the collateral of property has a lower current price. If there is an insolvent creditor, banks have a higher loss due to mortgages provided on the overvalued housing.[i]

Contribution

I would like to conduct empirical analysis based on existing data. Moreover, the literature has to be updated since statistical offices publish new figures each quartile. Most of the papers, to my knowledge, are not up to date. The increase of prices has changed last years. Thus, the expectation of raising interest rates by the Central Bank could slow down the growth. Results might be used as a warning sign for the real estate market. The model might determine whether the price of a currently selling asset is adequate or not. From another point of view, there are only few scholars in our country who have studied this economic area. For instance, I would like to build on the literature of M. Hlaváček, L. Komárek or P. Zemčík.

[i] E. Tham, "Ghost collateral' haunts loans across China's debt-laden banking system', Reuters, 2017. Available at: www.reuters.com: China Collateral Fake

Methodology

I am going to use data provided by the Czech statistical office and by Eurostat. Using those time series data, e.g. Housing index prices, the unemployment rate or the economic growth in a region, I would like to create a statistical model. Comparing relevant variables, the hypothesis of bubbling market should be tested. Moreover, applying the basic econometric methods, I would like to analyse the real estate market in the Czech Republic.

Outline

1. Introduction
2. Literature Review
3. Summarization of methods used in the analysis
4. Description of Collected Data
5. Analysis of the Model and Testing Hypothesis
6. Empirical macroeconomic consequences and their impact on the market
7. Conclusion

List of Academic Literature

- P. Poseděl, M. Vízek (2009). "House price determinants in transition and EU-15 countries". *Post-Communist Economies*, In: 21.3, pp. 327-343.
- M. D. Bordo, O. Jeanne (2002). "Boom-busts asset prices, economic instability, and monetary policy". *National bureau of economic research*,p 38.
- M. Hlaváček, L. Komárek (2009) "Housing Price Bubbles and their Determinants in the Czech Republic and its Regions". *Czech National Bank*, p. 54.
- P. Zemčík (2011). "Is There a Real Estate Bubble in the Czech Republic?". *Czech Journal of Economics and Finance*. In: 61.1, pp. 49-66.
- M. Hlaváček, L. Komárek (2011). "Regional Analysis of Housing Price Bubbles and Their Determinants in the Czech Republic". *Czech Journal of Economics and Finance*. In: 61.1, pp. 67-91, 2011.
- I. Kubicová, L. Komárek (2011). "The Classification and Identification of Assets Price Bubbles". *Czech Journal of Economics and Finance*. In: 61.1, pp. 34-48.
- S. Gilchrist and J. V. Leahy (2002). "Monetary policy and asset prices". *Journal of Monetary Economics*, p. 23.
- S. G. Cecchetti, H. Genberg, J. Lipsky, and S. Wadhwani (2000). "Asset Prices and Central Bank Policy". *The Geneva Report on the world Economy*. No. 2, p. 152.
- E. F. Fama (1970). "Efficient Capital Markets: A Review of Theory and Empirical Work". *The Journal of Finance*. In: 25.2, pp. 383-417.
- W. C. Hunter, G. G. Kaufman, and M. Pomerleano (2005). "Asset Price Bubbles: The Implications for Monetary". *Regulatory, and International Policies MIT Press*, pp. 3-80.

1 Introduction

Since 2016, the housing price in the Czech Republic has been increasing. The OECD analytical housing price indicators show the average growth of 7.2%, 11.7%, and 8.6% in the years 2016, 2017, and 2018, respectively. In Prague, the real price growth already started in 2013 (see Figure 1.1). Besides the years 2018 and 2017, however, this price growth is not as fast as in the years 2002/2003 and 2007/2008. During these two periods, there were even the property price bubbles which brought up a negative effect on the economy and the own housing was becoming less affordable for many people (Hlaváček and Komárek 2011; Zemčík 2011).

The price bubble would be even more harmful to the Czech economy where is traditionally the majority of people in own-occupied and privately-owned apartments, respectively.[ii] Even though rapid growth was partially given by "catch up" effect of the Czech economy after long decades of communism (Égert and Dubravko 2008). It is important to study the growth and decline of the housing price and their determinants since it is crucial for banks, investors, government, house-owners and last but not least for regulatory institutions. Furthermore, owner-occupied housing is the most significant part of many household's wealth, and the value of their property has a considerable impact on their consumption and savings (Case et al. 2004).

The other exception of the housing market from conventional markets of goods and services is the inelasticity of housing supply (Selim 2008). Thus, it is in individual's and government's concerns to avoid the property overpricing which would last more extended time period due to inelasticity.[iii] The regulatory institution in Czechia, which is interested in controlling the housing price fluctuation, is the Czech National Bank (CNB). The CNB financial stability report evaluates the sustainability of real estate prices on an annual basis. The CNB evaluates the price development by using the macroeconomics tools and calculates the overpricing or deviation from the fundamental pricing (Plašil and Andrle 2019). Therefore, I am going to search for the best predictors of property prices in the Czech Republic.

[ii] In 2014, there were 78.9% of households in privately-owned apartments in the Czech Republic. The EU(28) average was 70% (MRDCR 2018).

[iii] Hlaváček and Komárek (2009a) claims that the the burst of price bubbles in housing market lasts on average 4 years and causes larger output losses than the burst of bubbles in stock market (avg. 1.5 years).

The goal of this thesis is to choose the econometric models for time series and panel data, which will most precisely predict the price of second-hand purchased apartments. I evaluate the regressions by "goodness of fit" statistics as R-Squared and Residuals Mean Squared Errors (RMSE). Besides standard ordinary least squares (OLS) estimates, I also apply the Least Absolute Shrinkage and Selection Operator (LASSO) on my final regressions in order to choose the essential determinants within the explanatory variables. In section 3, I clearly state the null hypothesis for all independent variables as their expected effects on apartment prices.

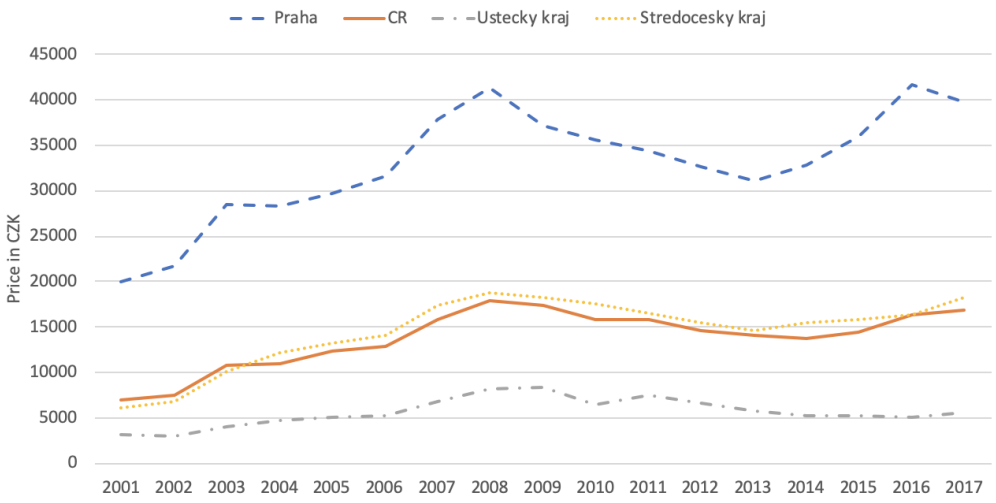
For this analysis, I will use the real prices of second-hand apartments provided by the Czech Statistical Office (CZSO). Thus, the price determinants for family houses or new apartments might be different. There will be used macroeconomic and demographic data, similarly as was for example used Hlaváček and Komárek (2009b). Besides the modern econometric methods applied on panel data set for the fourteen Czech regions, I will use the aggregated time series data set for the whole Czech Republic.

The unquestionable advantage of this thesis is that I can use a longer timeline, which is often essential for good predictive power (Hlaváček and Komárek 2009a; Plašil and Andrle 2019). We have to take into account the other researches on a similar topic from the Czech environment which are discussed in section 2 (see for instance Mikhed and Zemčík 2007; Hlaváček and Komárek 2009a; Zemčík 2011; Hlaváček and Komárek 2011). The aim is to extend these works. Nevertheless, I use only collected data from the CZSO and the CNB. Thus, there has to be a detached view in comparing results with these previous works since only slightly various data can provide very different results (Coskun et al. 2017).

Moreover, for the first time, there will be used Penalised Least Squares (PLS) method for the Czech housing data. This is a valuable contribution to the Czech academy. Generally, PLS method aims to decrease the forecast variance by shrinking the parameter estimates in the linear regression close or equal to zero. Thus, shrunk variables are "removed" from the set of explanatory variables (Smeekes and Wijler 2018).

As mentioned, I apply lasso regression since lasso deals with multicollinearity between explanatory variables which commonly occurs in the multiple regression (see the discussed problem of multicollinearity in these articles Kraha et al. 2012; Xin and Khalid 2018). Besides the comparison of the lasso regression with the ordinary least squares regression as the main goal of these two mentioned works, I compare more methods with lasso as the fixed effects, the first difference, and more time series alternatives. The lasso regression is discussed in section 4.

Figure 1.1: Real prices of second-hand apartment prices (100=2000)



Source: CZSO

The price determinants are not necessarily the same for every country, nor even for all regions within the country (Coskun et al. 2017). As Figure 1.1 suggests, Prague has a much higher real second-hand purchased apartment price than the rest of the republic since Prague is a region, itself. Moreover, price volatility is more extreme in Prague. It is reasonable that Prague may have different price determinants. However, I assume that the higher price in Prague is fundamentally explained as there are the highest values of the average wage, GDP per capita, or foreign direct investments (FDI). For this reason and a better interpretation, I run the regressions only for the whole Czech Republic using the time series and panel data.

This bachelor thesis is organised as follows. In section 2, there is a literature review of some related works on this topic. In section 3, data description and summary statistics take place. Not only the sources are discussed, but also the null hypothesis is stated for all explanatory variables, and they are described in detail. In section 4, I introduce the methodology and the structure of empirical models. In section 5, I present and comment on the results of the chosen regressions. Moreover, I highlight the most accurate regressions and their most important predictors. Lastly, the conclusion shortly summarises achievements and the fulfilment listed in section 1.

2 Literature Review

Several studies were done in the housing market around the world. Most of them analysed prices of real estate and thus, they are often from the periods when there was occurring some price discrepancy. For instance, (see Hlaváček and Komárek 2009b; Hlaváček and Komárek 2009a; Čadil 2009; Hlaváček and Komárek 2011; Zemčík 2011) were solving a rapid price growth in the first decade of the 21st century in the Czech Republic, when there was as well a fast growth in North America and Europe.

The focus was primarily on identifying price bubbles which are generally important for policymakers or for investors, who want to excess their returns. When the bubble bursts, the economy can be harmfully affected. Housing price bubbles are a threaten to the market. The most important factors of the Great Recession after the year 2007 were subprime mortgages and housing price growth (Calomiris, Longhofer, and Miles 2008).

Thus, regulatory institutions note and prevent the bubbles.[iv] How is actually the price bubble defined? An asset price bubble is defined as "an explosive and asymmetric deviation of the market price of an asset from its fundamental value" (Komárek and Kubicová 2011). The real estate bubble is usually defined as a discrepancy between property prices and fundamental property values (Zemčík 2011).

Identifying these appropriate fundamental property values is the goal of this thesis. Hlaváček and Komárek (2009b) searched for housing price determinants in the Czech market, which they divided into demand and supply factors. There are more demand factors, the major one being the disposable income of households, availability of housing loans, and unemployment rate. They also showed that some minor factors as divorce rate are statistically significant in determining the price of apartments.

A reasonable hypothesis that explains the significance of demographic factors is that after divorce, one household is split into two households. Two analyses were run depending on whether Prague was included or excluded. Nevertheless, the results remained almost unchanged, particularly the significance of coefficient estimates. Therefore, Prague should not be excluded since the macro-economic data fundamentally explain higher prices of apartments in Prague. For instance, the average real wage in 2004 in Prague is approximately equal to the average real wage in Czechia in 2017 or the yearly unemployment rate has not been higher than 4.2% in Prague.

[iv] For instance, see how CNB can regulate the boom in the housing market available at www.cnb.cz: Hrozbu nemovitostní bubliny umíme zkrotit.

Hlaváček and Komárek (2009a) continued in their previous research and attempted to determine housing bubbles. Real prices were used instead of nominal prices at this time. In order to identify the equilibrium path of the Czech Republic, they used price-to-income, price-to-rent ratio, and the Hodrick-Prescott filter. Čadil (2009) used, besides price-to-income ratio, VAR analysis to identify the possibility of a bubble on the Czech housing market for the period 1998-2006 for both apartment and family house prices. He also claims that flats are affected easier than family houses since there is a more straightforward speculative reason. Thus, determinants of apartment prices seem to be more important than, for instance family house prices, even though they might be very similar.

Others also used price determinants for identifying bubbles as Zemčík (2011) who applied the present value model and panel Granger causality techniques on prices and rents. He did not use regions for his study but 55 largest towns in Czechia. He also compared the supply-price based index (IRI)[v]and transaction-based index (CZSO). Talking about the potential price bubbles was important since prices of housing for the usual size of 68 m squares increased three times from 2001 until 2008. The dramatical increase was in Prague, where the most expensive apartments are located. Nevertheless, the results showed that there was only a small overpricing in the Czech Republic compared to the USA where was an evident and severe price bubble (Mikhed and Zemčík 2007).

All mentioned authors above agreed that there were two price bubbles between the end of the nineties and 2009, precisely, in the years 2002/2003 and 2007/2008, whereas the second bubble was smaller since more macro-economic and demographic explanations. On the other hand, the first bubble would be explained by a "catch-up" effect since the real estate was likely to be undervalued (Hlaváček and Komárek 2009b). All analyses also claim that speculations created the 2002/2003 bubble before the Czech Republic entered the EU and then stopped because speculative demand was in advance.

"Catching-up" effect fact also occurred in other countries of Central Europe, especially between the years 2002 and 2006 (Égert and Dubravko 2008). They applied a dynamic OLS regression on data from the OECD. They searched for the determinants of house prices in eight transition economies of Central and Eastern Europe (CEE) and 19 OECD countries. "The catching-up" effect was confirmed in their work since the house prices grew two times faster in CEE from 2002 to 2006 than in the countries of OECD. They also claim that the rapid credit growth in CEE should have a lower impact on house price growth than in more developed OECD countries.

[v]Data from the Institute of Regional Information at Brno were used.

Restrictions placed on rents partly give an understanding of real estate in Czechia. The Czech housing market was still partly restricted in the first decade of the 21st century, and a share of mortgages in the country was not as risky as in the USA or in Great Britain.[vi]Therefore, overvaluation of the housing market in Czechia in 2002/2003 and 2007/2008 was lower than in more opened market countries (Čadil 2009).

Besides fundamental property values, what are the other techniques for pointing out price bubbles? Mikhed and Zemčík (2007) investigated bubbles based on the price-rent ratio. The statistical discrepancy was tested by stationarity. "If the prices are non-stationary, but rents are not, we view that as an indicator of a bubble"[vii], they state in their paper. In other words, if the average price has been growing over time and the growth of rent does not support this growth, then we can have a suspicion on the bubble, and we cannot explain the irrational demand for declining return on investments in real estate fundamentally. The fundamental explanatory variable is the rent price in this case.

The similar approach is the price-to-income ratio where the income plays the same role as the rent. There is shown that Prague had already, between the years 2000 and 2009, the highest price-to-income ratio among other large cities in the Czech Republic (Hlaváček and Komárek 2011). This claim is valid even if we compare Prague with other large cities in Europe. Prague had by more than 30% the higher actual price per metre-square of an apartment than its fitted value in 2012. This result was the same for the Ordinary Least Squares and quantile regressions. It ranked Prague on the third place behind London and Rome in the most "overvalued" cities in Europe, according to the internet offer price for flats (Kholodilin 2012).

This moves us to the term "housing affordability", which is strongly connected to housing price determinants. The monthly average income is supposed to be significant for determining of housing prices since housing affordability is calculated by using this income variable and housing prices (See for instance Stone 2006; Kostelecký and Vobecká 2009; Stone 2010; Bramley 2011). Thus, if the ratio income - housing price is constant, then the housing affordability should be constant as well. This equality is essential again for regulatory institutions since housing price is strongly pro-cyclical and tends to grow faster than other macroeconomic variables.

[vi]See more about price regulations posted by the Czech Ministry of Finance available at www.mfcr.cz: Vývoj cenové regulace v jednotlivých odvětví.

[vii]See the summary of the paper, p.24.

[vii]See again the interview from CNB websites available at www.cnb.cz: Hrozbu nemovitostní bubliny umíme zkrotit.

Papers, mentioned in the paragraph above, usually discuss housing affordability in relation to poverty. They would like to answer the question: is the price chosen appropriately? Cai and Lu (2015) claim that there are some factors which rise or decrease the price. For instance, better accessibility increases the price of housing. It means that the housing is located nearby a developed infrastructure with good public transport. This work dealt with a trade-off between affordability, accessibility, amenity, and adequacy. The amenity is the condition of a housing unit which I do not take into account in this thesis since it cannot be measured on the base of quantitative data from the CZSO.

However, adequacy is searched in the thesis since it measures how the housing unit should be valued. Affordability restricts mostly the poorest cast of the population. This case study from China showed that income-constrained consumers have been facing unaffordable housing since their houses and apartments are usually overvalued according to these four factors. Moreover, they have been spending the highest portion of their income on housing. The adequate portion of income which should be spent on housing is considered 25-30%, in some countries 40% (see also Thalmann 1999; Yates and Gabriel 2006).

More works were done in emerging markets which attract attention due to their usual fast economic growth.[viii]Coskun et al. (2017) run the research of bubble risk in Turkey with the use of Kalman filter and time series analysis – OLS, ARMA models. They defined a bubble based on actual and fundamental house price interactions. According to this paper, there are three possibilities. If the actual price (P_t) is higher than the fundamental price (P_t^f)), then there has to be some bubble component. In the opposite case, where $P_t < P_t^f$), housing price is undervalued, otherwise the housing market is in its equilibrium.

Based on these results, people decide whether to hold, sell, or buy real estate. The housing market in Turkey was only slightly overvalued between 2010 and 2014, but Coskun et al. (2017) claim that the growing price of real estate is typical for emerging markets. However, they did not use as many explanatory variables as (Hlaváček and Komárek 2009b), for example, average wage and demographic variables. The main goal of this thesis is not to search for the price bubbles, but how well the variables explain the real property prices. In other words, do our models predict the actual price, or do they suggest price misalignment.

[viii]See the results from the research on the emerging economies done by the McKinsey Global Institute available at www.mckinsey.com: Outperformers High Growth Emerging Economies and the Companies that Propel them.

Another approach how to determine house prices is a hedonic model that evaluates the effect of characteristics of real estates on their prices, where the term "hedonic" expresses the relative importance of various components among others. For this model, there are needed a detailed data of particular offered properties besides a location and size as the type of building, rooms, or heating (Selim 2008). For example, Selim (2008) shows in his results that the prices of houses in the urban area in category 0-5 years are supposed to be by 8% higher than 5-10 years old houses.

In this thesis, there are used Penalised Least Squares Estimates, more precisely lasso method, firstly mentioned and described by Tibshirani (1996). There are also papers which applies lasso on a similar topic. For instance, Xin and Khalid (2018) studied house price modelling. They claim that in most cases there is multicollinearity between explanatory variables in the housing market models. Similarly, as Selim (2008), they used qualitative housing data, for example commercial location and heating quality for determining a sale price, but they applied ridge and lasso regression. Finally, they found that lasso regression had better results than ridge regression for their data due to the lower Mean Square Error and higher adjusted R-squared.

Using the lasso method is useful for a trade-off: low variance against low bias. Chan-Lau (2017) summarised that the lasso tends to outperform traditional statistical models which deal with stress tests. "The stress test specifies a large set of primary explanatory variables for which there are only a few observation" (Chan-Lau 2017). Lasso method will be more described in section 4 and section 3 outlines the explanatory variables for my models. Thus, there will be shown that it is convenient to use lasso regression for shrinking the coefficients and reducing the Mean Squared Error of the models.

There are other studies which describe the use of penalised regression methods for forecasting (see Li and Chen 2014; Smeekes and Wijler 2018). They state that it is convenient to use lasso when the total number of predictors is substantial since the classical ordinary least squares estimators usually have a large variance. Therefore, the fitted model is unstable in predicting over time. Moreover, when the frequency of data is low like in this thesis (two data sets with quarterly and yearly frequency), the over-fitting is more likely to occur. Lasso method should identify the subset of zero coefficients that can be excluded from the model, and my model should obtain estimators with a lower variance. Similarly as Chan-Lau (2017), Smeekes and Wijler (2018) test if the forecasts are close to the actual value by using the RMSE on the out-of-sample. Comparing the prediction of dynamic factoring model (DFM) with PLS estimators, lasso and ridge are significantly better than DFM. Furthermore, the lasso approach has a greater statistical gain in forecasting many variables than ridge regression.

3 Data Description

3.1 Data Structure

In this thesis, there are used two different data sets which contain macroeconomic, monetary and demographic data from the Czech Statistical Office (CZSO) and the Czech National Bank (CNB). Firstly, I use panel data which are created from 14 Czech regions. Panel data are collected annually between the years 2002 and 2017. Secondly, the times series data is created from aggregate country quarterly data between the years 2000 and 2018.

In order to synchronise data, some variables are only used in annual periods, thus only for panel data models (see Table 3.2). Some important macroeconomic data, such as the inflation rate or the unemployment rate, are collected monthly or quarterly, while other variables that we need for the analysis, as the divorce rate, are available only on an annual basis. Therefore, I use the other sample of explanatory variables for time series analysis (see Table 3.3).

The cooperation between the CZSO and the Ministry of Finance started at the beginning of 1998. Since the period restricts the maximal sample, the analysis can be from this date onward, only. Moreover, not all the data we need are available from 1998, and I have had use panel data from 2002 to 2017 only. The regional property prices were not available for 2018 since the Ministry of Finance has been delayed sharing documents with the CZSO. Data for a previous year are usually available at the end of the current year for the different regions (CZSO 2018). Whereas, the CZSO provides a better aggregated data collection for the whole Czech Republic. Thus, quarterly time series data are available from 2000 to 2018.

The Ministry of Finance collects data on property transfer tax returns. This approach has cons and pros. The advantage is the completeness of data which means that we have involved all the second-purchased apartments in our data. On the other hand, the release of property transfer tax data has been delayed. Moreover, as mentioned, this data only concerns secondhand housing since new houses are not subject to property transfer tax.[ix] The subjects to property transfer tax are as follows: Family houses, apartments, residential houses, garages, and building plots.

[ix] Hypothetically, there might be rare cases when a developer sells an apartment unit to a "speculator" who will sell the apartment to the final owner afterward. Then, we might have some exceptions when there is a new apartment involved in our database.

The essential disadvantage that I observed are inconsistent samples for each year (apartments from different area, larger/smaller samples). The CZSO tries to eliminate these statistical deviations but it is not possible in all cases. We can see in Figure 1.1 that the real apartment price declined in 2017. Even though the other sources claim that the house price index in Czechia grew by 11.7% (OECD), 8.9% (CZSO), or in Prague 7.7% (CZSO), which is higher growth than the inflation rate in 2017.[x] Thus, I assume that the decrease of real second-hand prices is given by a "cheaper" sample given by purchasing less expensive apartments in suburbs.

Table 3.1: Explanation of Variables

Dependent variable: Apartment price						
	Unit	Model	Frequency	Side	Source	$^{*}Corr$
Apartment price	CZK per m^2	TS, PD	Quarter, Annual	-	CZSO	1
Building plot price	CZK per m^2	TS	Quarter	Supply	CZSO	+
Finished apartments	No. of units	TS,PD	Quarter, Annual	Supply	CZSO	-
Started apartments	No. of units	TS,PD	Quarter, Annual	Supply	CZSO	+
Average wage	CZK per person	TS,PD	Quarter, Annual	Demand	CZSO	+
GDP per capita	CZK per person	TS,PD	Quarter, Annual	Demand	CZSO	+
Unemployment rate	Avg. percentage	TS,PD	Quarter, Annual	Demand	CZSO	-
Population density	No. of people per km^2	PD	Annual	Demand	CZSO	+
Net migration	No. of people	PD	Annual	Demand	CZSO	+
Population growth	No. of new-born	PD	Annual	Demand	CZSO	+/-
Marriage rate	No. of marriages	PD	Annual	Demand	CZSO	+/-
Divorce rate	No. of divorces	PD	Annual	Demand	CZSO	+
Interest rate	Percentage	TS, PD	Quarter, Annual	Demand	CNB	-
FDI	CZK per person	PD	Annual	Demand	CNB	+

Note: $^{*}Corr$ is assumed to be a sign of correlation between an independent variable and *Apartment price* (a sign of coefficient estimator in the model).

[x] See the OECD data available at https://stats.oecd.org. See the CZSO data available at https://www.czso.cz.

3.2 Hypothesis Statement

The last column of Table 3.1 says the expected sign of the coefficient. In other words, what the relationship is between a particular explanatory variable and the dependent variable. The column Corr is a "summary" of hypotheses which I state in subsection 3.4, and subsection 3.5. The hypotheses are stated informally, as the expected influence of an explanatory variable on the response variable.

The informal discussion states hypotheses of the following forms.
Firstly, the null hypothesis:

$$H_0 : \beta_i = 0, \tag{3.1}$$

where β_i is a coefficient of an ith explanatory variable, and H_0 means: "The explanatory variable has no effect on the dependent variable."
Secondly, the alternative hypothesis:

$$H_1 : \beta_i \neq 0, \tag{3.2}$$

where β_i is a coefficient of an ith explanatory variable, and H_1 means: "The explanatory variable has an effect on the dependent variable." Whereas, if β_i has the expected sign, I reject the null hypothesis in favour of my expectation. On the other hand, if β_i has the unexpected sign, I reject the null hypothesis in disfavour with my expectation. The following subsection describes the response variable in detail.

3.3 Apartment Prices

As a dependent variable, I have chosen apartment prices with the base year 2000 for the analysis. More precisely, it is the average apartment price for a given region and period in real terms. Why should be apartments convenient for this analysis? Firstly, we have chosen apartments since the average apartment's wear is almost equal among regions.[xi] In other words, most of the purchased apartments are under similar conditions. Moreover, apartments have a high homogeneity in the sample, which is far the best result comparing to other types of real estate (CZSO 2018).

Secondly, the homogeneity takes us closer to the fact that the price mostly depends on the time period and also on the region. This is important for the macroeconomic data since I cannot regress a hedonic model and I can use aggregated data. Lastly, the CZSO data shows that the average size of apartments across regions is almost equal, between 60 m^2 *and* 68 m^2.[xii] More precisely, the last housing census in 2011 showed that the average total area per inhabited dwellings in multi-dwelling building was 68.5 m^2 (MRDCR 2018).

[xi] More information about the methodology of property wear available at https://www.czso.cz.
[xii] When we talk about the average apartment, we mean 68 metres squares 2 bedroom flat.

Hypothetically, I have avoided diminishing prices per meter square with a larger apartment. A similar approach was used in the paper from Kholodilin (2012), who analysed internet offer price determinants across the European cities. Western Europe has larger average apartments than the central and eastern Europe (Kholodilin 2012). Thus, Czech apartments are relatively homogenous and suitable for the analysis.

In the following subsections, I have described the variables which are used in the models with more details than it is shown in Table 3.1, Table 3.2, and Table 3.3. There are stated hypothesis about the expected influences of particular variables on apartment prices which summarizes Table 3.1. Similarly as Hlaváček and Komárek (2009b), explanatory variables will be said to be a subset of demand or supply side. The important note is that all the prices are in real terms with the base year of 2000 (100=2000).

Thus, the inflation rate is not used as an explanatory variable. As Hlaváček and Komárek (2011) claim, models with real values have more robust estimators. In addition, housing prices are a part of the inflation calculations, and also for this reason, it is better to use inflation rate only for the calculation of real values. Moreover, there would be a spurious growth between the nominal apartment price and the average wage.

3.4 Supply Side Variables

The supply side contains factors which are affected by the expectation of the present value of builder's profit from selling a house (Poterba 1984). Thus, the supply of new apartments depends on the current housing prices and other predictions of developers, whereas the supply of existing housing is inelastic.

Building Plot Price

The building plot price is used only for the time series analysis of the Czech Republic. Region data are not available in the CZSO database. It is expected that there is a positive correlation between *apartment price* and *building plot price*.

Finished Apartments

Finished apartments are considered to be apartments that can be placed on the market. Thus, I expect that higher apartment saturation increases supply and lowers the price. We take into account only housing units that are built in an apartment building. This variable might be endogenous in our model. One can easily imagine it is determined simultaneously with the apartment prices. Therefore, there is a reverse causality in our model, and the strict exogeneity assumption is violated.

Nevertheless, if we assume that finished housing units for a given year depend only on the past prices, then finished housing units is an exogenous variable to the current prices. Thus, *finished apartments*$_t$ affects *apartment price*$_t$, but *apartment price*$_t$ does not affect *finished apartments*$_t$. The current prices are linked to the past prices, via this channel, they are linked to the current finished housing units.

Started Apartments

Started apartments are considered apartments that received a building permit in the particular period. We take into account only housing units that are supposed to be built in an apartment building. This variable is assumed to be endogenous since the number of building operations is likely to react to the actual housing price. In the case, sharper price growth, the higher number of started apartments is expected, and vice versa.

3.5 Demand Side Variables

The demand side contains factors which are affected by consumer needs. These macroeconomic, demographic and monetary variables are listed below.

Average Wage

The average real wages are used since our dependent variable is also determined in real prices, and I expect a positive correlation between these two variables. I use average monthly wage in both data sets. The base year is 2000. The average real wage is an amount of money before taxation, and it expresses consumer's purchasing power more precisely than the nominal wage. I cannot use disposable income since there were a different tax rates on income in the past. Moreover, there are many tax allowances in Czechia.

GDP per Capita

We have data for both regions and countries. A higher GDP per capita growth might encourage people to get divorced, get married, or have children. For example, Hellerstein and Morrill (2011) shows that divorce rate was slightly higher in times of economic booms than during recessions in the USA between the years 1976 and 2009. A region with a high GDP per capita could be attractive for migration. Consequently, a need for more housing units is expected.

Unemployment Rate

"Unemployment rate is derived as the ratio of the number of job applicants out of work to the number of employment."[xiii] Higher unemployment implies less disposable income of households and people have less money for investing in property. Regions with higher unemployment are also not popular as a migration destination since people might be terrified of losing their job and the criminality might be higher as well, which partly claim some studies (see for example Phillips and Land 2012; Fallahi and Rodríguez 2014).

[xiii] Source: www.czso.cz: Unemployment Rate.

Population Density

There are used own simple calculations where the number of inhabitants for each year is divided by region area. This variable is used only for panel data. It is assumed that higher *population density* could increase the price of the apartments since there should be a higher demand for housing. However, a higher population density should imply a higher housing density. Thus, this variable could be without any effect. Despite this fact, population density can be considered as a "dummy variable" for Prague since Prague has a significantly higher population density than the other regions.

Net migration

Net migration equals to the number of incomers minus the number of moved out people in the region. I use *net migration* and all the following demographic factors for panel data only. More incomers imply a higher demand for housing in the region. Thus, the increase in the apartment's prices is expected.

Population Growth

Population growth means the number of newborn children minus the number of deceased people in a given year. A higher number of born children might cause the needs of families to seek for a larger housing. On the contrary, they will likely place their previous housing at the market, and there could be another offered apartments that belonged to deceased people.

Marriage Rate

The *marriage rate* is the number of marriages in a given region per year. *Marriage rate* occurred to be a significant explanatory variable in the paper from Hlaváček and Komárek (2009b). They expected a positive sign of *marriage rate* since a new married couple will search for new housing. On the other hand, one might say that it depends on whether they had before the wedding an own apartment (negative sign expected), already shared a household (no change) or they could live with their parents or roommates (positive sign expected).

Divorce Rate

The *divorce rate* is the number of divorces in a given region per year. A higher divorce rate, more split households, and consequently, higher demand for housing is expected. Hlaváček and Komárek (2009a) showed that the *divorce rate* is a significant demographic variable in determining apartment prices.

Interest Rate

The *interest rate* is a repo rate which is set by the Czech Central Bank. *Interest rate* should be involved since this rate affects the price of mortgages, consequently, the number of mortgages. Similarly, as *finished apartments*, this variable might also be endogenous. One may say that the Czech Central bank changes a discount rate according to the growth of property prices. The central bank usually increases the interest rate on mortgages when there is a rapid growth in property prices. See Figure 6.15 where is shown the housing price index growth and interest rate change between the years 2000 and 2018.

The interest rate for mortgages follows year to year property price growth.[xiv] According to Figure 6.15, one may expect that *interest rate* is negatively correlated with *apartment price* since a decrease of interest probably means the increase of apartment prices. However, this link is delayed. Thus, the correlation could also be positive.

Foreign Direct Investments

Foreign direct investments (FDI) are the total foreign direct investments into the regions from abroad divided by the regional population. Direct investment = equity capital + reinvested earnings + other capital.[xv]I would have prefered to use direct investments only into the property, but the CNB does not provide the required data for regions for a longer timeline.

Higher FDI, the higher attractiveness of a region is assumed, and consequently, I expect an increase in apartment prices. Table 3.3 shows a high standard deviation in FDI. Prague has the highest FDI per person. The high standard deviation means that some regions are unattractive for foreign investors comparing to Prague.

[xiv] See how the CNB has changed two-week repo rate over time available at https://www.cnb.cz: How was the CNB Two Week Repo Rate Changed.

[xv] Czech National Bank takes over the definition of FDI from OECD:

"Foreign direct investment reflects the objective of obtaining a lasting interest by a resident entity in one economy ("direct investor") in an entity resident in an economy other than that of the investor ("direct investment enterprise")...Direct investment involves both the initial transaction between the two entities and all subsequent capital transactions between them and among affiliated enterprises, both incorporated and unincorporated."

3.6 Summary of Descriptive Statistics

Table 3.2: Descriptive Statistics - Panel Data

Statistic	N	Mean	Min	Pctl(25)	Pctl(75)	Max	St. Dev.
Apartment price	238	12,967	2,909	8,900	14,782	41,670	7,041
Finished apartments	252	723.2	4	151.8	761.8	7,908	1,080.8
Started apartments	252	726	0	161	826.8	6,010	1,030.2
Average wage	252	15,914	11,220	14,239	17,062	25,177	2,619
GDP per capita	252	331,869	178,346	260,338	349,744	997,560	135,945
Unemployment rate	252	6.575	1.900	4.600	8.025	16.000	2.820
Population density	252	291.4	62.1	91.8	155.5	2,656	610.6
Net migration	224	1,960.9	−5,297	−408.5	2,034.3	25,873	4,653.7
Population growth	224	47.7	−3,643	−499.8	522.5	3,125	1,033.7
Marriage rate	224	3,519.4	287	2,402.5	5,212.3	7,149	1,612.6
Divorce rate	224	2,084.8	734	1,349	2,954.5	4,404	967.3
Interest rate	252	1.764	0.250	0.250	2.500	5.250	1.534
FDI	224	189,845	35,643	91,027	196,601	1,386,016	230,666

Note: Annual data 2000 - 2017. Source: CZSO, CNB.

Table 3.3: Descriptive Statistics - Time Series Data

Statistic	N	Mean	Min	Pctl(25)	Pctl(75)	Max	St. Dev.
Apartment price	76	13,922	8,400	12,529	15,264	18,920	2,553
Building plot price	76	1,235	618.6	1,080.4	1,419.4	1,949.4	297.1
Finished apartments	76	2,310.3	814	1,588.2	2,851.5	9,085	1,185.8
Started apartments	76	2,201.6	531	1,319.2	2,887.8	4,903	1,040.8
Average wage	76	16,430	11,593	14,870	17,778	21,105	2,210
GDP per capita	76	72,008	52,354	66,473	77,654	90,299	8,336
Unemployment rate	76	6.309	2.040	4.907	7.772	9.510	1.852
Interest rate	76	1.776	0.050	0.200	2.500	5.250	1.630

Note: Quarter data 2000 - 2018. Source: CZSO, CNB.

4 Methodology and Empirical Models

In this section, econometric and statistical methods are described in detail. I discuss the assumptions, pros, and cons of the models. In order to analyse the data, I use panel data created from 14 different Czech regions and aggregated time series data for the whole country. It is important to take into account that I work with two data sets. In the first part, I describe models for panel data. In the second part, the models for time series are described. Lastly, I show how to apply penalised least squares methods on the models with the best performance, and I examine why the use of PLS is convenient for our data, particularly, why lasso regression should improve predictive power.

4.1 Panel data approach

Apartment price determination is a complex process, and I can hardly claim that we control for all significant variables. Omitted variables could cause bias and we would be likely to obtain biased and inconsistent estimators (Wooldridge 2012).

Ordinary least squares regression

Panel data contains many specific time-constant variables that differ across regions. For example, Prague is expected to be a particular case since it is by far the largest city in the Czech Republic, being more cosmopolitan than the rest of the country. Primarily, Prague is a region by itself whereas other regions have their local capitals. Therefore, specific time-constant variables would cause unobserved heterogeneity, which is not fixed by the classical OLS model. If we had used only a few periods, it would be controlled by adding dummy variables for these periods. However, the timeline has 16 periods. It would be tough for interpretation to add dummy variables for all years (Wooldridge 2012).

Dummy variables regression

For this reason, I tried to add dummy variables only for the most extreme regions - Prague (the highest prices) and Ustecky region (the lowest prices) and binary variables for the years when the apartment prices had the sharpest growth in prices - the year 2002 and 2007. Nevertheless, the results were very unrealistic for the interpretation with the Prague coefficient -22750 CZK. I expected this coefficient to be positive. The negative coefficient would only mean that apartments are overpriced in Prague. However, this model does not give us an accurate prediction. Thus, I will not use a dummy variable regression in further analysis neither, and there will be discussed more advanced econometric methods - random effects, fixed effects, and first difference.

Random effects regression

Random effects (RE) regression is usually used to increase the efficiency of coefficients prediction when there are constant explanatory variables over time. However, RE assumes that the unobserved composite error is uncorrelated with all explanatory variables as it is stated in Equation 4.1 (Wooldridge 2012). In other words, strict exogeneity should be satisfied. However, strict exogeneity is impossible since the specific time-constant variable has to be correlated to the region's data.

Therefore, we should search for an alternative to RE. The Hausman test, introduced by Hausman (1978), indicates whether there is some statistically significant difference between random and fixed effects estimators or not. If we cannot reject the null hypothesis stated in Equation 4.1, then it does not matter whether we use RE or FE. In other words, the Hausman test is for this hypothesis:

$$\begin{aligned} H_0 &: y_{it} = \beta_1 x_{it_1} + \beta_0 + \ldots\ldots + \beta_k x_{it_k} + a_i + u_{it} \\ &\quad where\ Cov\left(x_{it_j}, a_i\right) = 0,\ for\ all\ t = 1, 2..., T\ and\ j = 1, 2, ..., k. \\ H_1 &: Cov\left(x_{it_j}, a_i\right) \neq 0 \end{aligned} \tag{4.1}$$

The resulting p-value from the Hausman test is lower than 0.05 in Table 4.1. In other words, the estimators of FE and RE are different at 5% significance level. As before, this would be caused by the time-constant unobservable characteristics of a particular region. Thus, we should not use random effects in further analysis. As the alternative, I use the fixed effects regression, which eliminates the region's constants and better treat the strict exogeneity assumption.

Table 4.1: Hausman test for panel data

	(Chi-squared)	(DF)	(p-value)
Results	12.797	12	0.0384

Alternative hypothesis: one model is inconsistent

Fixed effects and First difference regressions

After rejecting the random effects, we must design our model in a way that is more likely avoiding inconsistent estimates of parameters. As an alternative, we can use fixed effects. Besides FE, the difference between two periods of the same region eliminates time-constant coefficients in the model as well. Thus, we can use the first difference or fixed effect regressions.

Firstly, if the no serial correlation assumption holds, we instead use the fixed effects model since using a first difference model would introduce the undesirable serial correlation. Secondly, if no serial correlation is violated, we prefer to use a first difference model since errors are correlated in the two different periods. The difference between the two following periods might fix the serial correlation (Wooldridge 2012). However, if there is still an essential negative serial correlation after the first difference in errors, fixed effects should be better according to Wooldridge (2012).

We test for the serial correlation in the autoregressive model of order q:

$$u_{it} = \rho_1 u_{it-1} + \rho_2 u_{it-2} + \ldots\ldots + \rho_q u_{it-q} + e_{it}, \tag{4.2}$$

with the null hypothesis of no serial correlation: $H_0 : \rho_1 = 0, \rho_2 = 0, \ldots\ldots, \rho_q = 0$, and Lagrange multiplier: $LM \sim \chi^2_q$, can be tested with the use of the Breusch-Godfrey (BG) test. The LM statistics for testing $H_0 : \rho_1 = 0, \rho_2 = 0, \ldots\ldots, \rho_q = 0$, has the following shape:

$$LM = (n - q)\, R^2_{\hat{u}}, \tag{4.3}$$

where $R^2_{\hat{u}}$ is R-squared from the regression $\hat{u}_t$ on $x_{t1}, x_{t2}, \ldots\ldots, x_{tk}$ (Wooldridge 2012).

We can reject no serial correlation in both cases at the low significance level (see Table 4.2. We have a serial correlation in fixed effects as well as in the first differences. Thus, we should use both models and discuss their results. In order to have robust estimates, I will use heteroskedasticity and autocorrelation consistent (HAC) standard errors to both models (as it is described in Zeileis 2004). See the summary table with results in Table 5.1.

Table 4.2: Breusch-Godfrey test for panel data

	(Chi-squared)	(DF)	(p-value)
First difference	58.427	16	0.0000009614
Fixed effects	87.109	16	0.000000000008498

Thus, I will use the fixed effect regression in the further analysis, similarly as Hlaváček and Komárek (2009a). The FE regression has the following form:

$$\begin{aligned} apartp_{it} - \overline{apartp_{it}} = {} & \beta_1 \left(finapart_{it} - \overline{finapart_{it}}\right) + \beta_2 \left(strapart_{it} - \overline{strapart_{it}}\right) \\ & + \beta_3 \left(wage_{it} - \overline{wage_{it}}\right) + \beta_4 \left(gdpcap_{it} - \overline{gdpcap_{it}}\right) \\ & + \beta_5 \left(popdens_{it} - \overline{popdens_{it}}\right) + \beta_6 \left(migr_{it} - \overline{migr_{it}}\right) \\ & + \beta_7 \left(popgrwth_{it} - \overline{popgrwth_{it}}\right) + \beta_8 \left(marriage_{it} - \overline{marriage_{it}}\right) \\ & + \beta_9 \left(divorce_{it} - \overline{divorce_{it}}\right) + \beta_{10} \left(intrat_{it} - \overline{intrat_{it}}\right) \\ & + \beta_{11} \left(fdi_{it} - \overline{fdi_{it}}\right) + \beta_{12} \left(unemp_{it} - \overline{unemp_{it}}\right) + \left(u_{it} - \overline{u_{it}}\right), \end{aligned} \tag{4.4}$$

where i = 1,.....,14 determines a region, t = 2002,.....,2017 stands for an year and u_{it} are unobserved variables.

Hlaváček and Komárek (2009a) also used the first difference, but only for chosen non-stationary variables. Otherwise, Hadri panel unit root test suggested them to use the FE, only. The BG test used in this thesis rejected that any regression is better regarding serial correlation assumption. Therefore, I use the first difference for all variables. The FD regression has the following form:

$$\begin{aligned} apartp_{it} - apartp_{it-1} = {} & \beta_1 \left(findapart_{it} - finapart_{it-1}\right) + \beta_2 \left(strdapart_{it} - strapart_{it-1}\right) \\ & + \beta_3 \left(avgwage_{it} - avgwage_{it-1}\right) + \beta_4 \left(gdpcap_{it} - gdpcap_{it-1}\right) \\ & + \beta_5 \left(popdens_{it} - popdens_{it-1}\right) + \beta_6 \left(migr_{it} - migr_{it-1}\right) \\ & + \beta_7 \left(popgrwth_{it} - popgrwth_{it-1}\right) + \beta_8 \left(marriage_{it} - marriage_{it-1}\right) \\ & + \beta_9 \left(divorce_{it} - divorce_{it-1}\right) + \beta_{10} \left(intrat_{it} - intrat_{it-1}\right) \\ & + \beta_{11} \left(fdi_{it} - fdi_{it-1}\right) + \beta_{12} \left(unemp_{it} - unemp_{it-1}\right) + \left(u_{it} - u_{it-1}\right), \end{aligned} \tag{4.5}$$

where i = 1,.....,14 determines a region, t = 2003,.....,2017 stands for an year and u_{it} are unobserved variables.

4.2 Time series approach

In the previous part, I discussed methods for determining the best estimates in panel data. In this subsection, I will describe the methods that are used in time series models. As mentioned, in order to have more observations, quarter data are used in time series for the whole Czech Republic.

OLS regression

Firstly, we test for stationarity. The unit root test is conducted for testing stationarity by using the autoregressive model of order one, known also as AR(1):

$$y_t = \rho y_{t-1} + \eta_t, \tag{4.6}$$

and

$$y_t = \rho y_{t-4} + \eta_t, \tag{4.7}$$

where the null hypothesis stands for $\rho = 1$, which says that our regression is non-stationary and the alternative hypothesis is given as $\rho < 1$. Looking at Table 4.3, we cannot reject the null hypothesis at 1% significance level, and thus, our regression is non-stationary (Hylleberg et al. 1990). Since we have quarter data, unit root test is done for the $period_{t-1}$(previous quarter) as well as for the $period_{t-4}$(the same quarter in the previous year) similarly as Hylleberg et al. (1990) who used Equation 4.6 and Equation 4.7 for testing stationarity at seasonal data.

Secondly, OLS model has to be tested for the serial correlation by using the AR(1) for residuals. In Table 4.4, there is again seen that there is a serial correlation for either, $residuals_{t-1}$ (previous quarter) or $residuals_{t-4}$ (the same quarter in the previous year). In both cases, there is a significant positive correlation between residuals in two different periods.

Table 4.3: Unit Root Test for Stationarity

	Dependent variable: Apartment price	
	$AR(1)_{t-1}$	$AR(1)_{t-4}$
Apartment $price_{t-1}$	0.991^{***}	
	(0.004)	
Apartment $price_{t-4}$		0.963^{***}
		(0.012)
Observations	75	72
R^2	0.999	0.989
Adjusted R^2	0.999	0.989
Residual Std. Error	436.182 (df = 74)	1,409.251 (df = 71)
F Statistic	$73{,}826.580^{***}$ (df = 1; 74)	$6{,}576.847^{***}$ (df = 1; 71)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 4.4: Serial Correlation of Time Series

	Dependent variable: Residuals$_t$	
	AR(1)$_{t-1}$	AR(1)$_{t-4}$
Residuals$_{t-1}$	0.535***	
	(0.094)	
Residuals$_{t-4}$		0.347***
		(0.107)
Observations	75	72
R^2	0.303	0.129
Adjusted R^2	0.294	0.116
Residual Std. Error	708.820 (df = 74)	785.392 (df = 71)
F Statistic	32.181*** (df = 1; 74)	10.470*** (df = 1; 71)
Note:		*p<0.1; **p<0.05; ***p<0.01

Thus, heteroskedasticity and autocorrelation consistent (HAC) standard errors should take place in order to have more efficient and robust estimators (again as described Zeileis 2004). Standard feasible GLS estimators cannot be used because strict exogeneity assumption does not hold. For instance, the classical TS OLS model was used in the analyses of (Hlaváček and Komárek 2009a; Čadil 2009; Coskun et al. 2017). The OLS model for time series has the following form:

$$\begin{aligned} apartp_t = \beta_0 + \beta_1 (finapart_t) + \beta_2 (strapart_t) + \beta_3 (avgwage_t) \\ + \beta_4 (gdpcap_t) + \beta_5 (unemp_t) + \beta_6 (bplotp_t) + \beta_7 (intrat_t) + u_t, \end{aligned} \tag{4.8}$$

where t = 2000Q1, 2000Q2,.....,2018Q4 determines a quarters in different years and u_t are unobserved variables.

Trending regression

We can consider more models for time series. I assume that some variables have a trend such as *average wage, GDP capita*, and especially *apartment price*. Therefore, the time trend can be added if it has a significant effect on the other estimators. In our case, the trend has a significant coefficient (see Table 5.2).

Adding the time trend might be helpful since some variables can be significant only for having a trend. In some cases, it could result in a spurious regression problem. We also eliminate the problem that unobserved trending factors affect the dependent variable. It might misrepresent the results. On the other hand, I have to note that the trend does not treat or violate our Gauss-Markov assumptions (Wooldridge 2012). Nevertheless, it could increase models predictive power, and I will use the following model for the final comparison:

$$\begin{aligned} apartp_t = {} & \beta_0 + \beta_1 \left(finapart_t\right) + \beta_2 \left(strapart_t\right) + \beta_3 \left(avgwage_t\right) \\ & + \beta_4 \left(gdpcap_t\right) + \beta_5 \left(unemp_t\right) + \beta_6 \left(bplotp_t\right) + \beta_7 \left(intrat_t\right) \\ & + \beta_8 (trend) + u_t, \end{aligned} \tag{4.9}$$

where t = 2000Q1, 2000Q2,.....,2018Q4 determines a quarters in different years and u_t are unobserved variables.

Seasonality regression

Quarter data might contain seasonality trend, even though that many available data are seasonally adjusted in advance. Seasonality can be avoided if we add dummy variables for different quarters (Wooldridge 2012). Seasonal adjustments and their applying on models are also discussed Hylleberg et al. (1990). Seasonality regression has the following form:

$$\begin{aligned} apartp_t = {} & \beta_0 + \beta_1 \left(finapart_t\right) + \beta_2 \left(strapart_t\right) + \beta_3 \left(avgwage_t\right) \\ & + \beta_4 \left(gdpcap_t\right) + \beta_5 \left(unemp_t\right) + \beta_6 \left(bplotp_t\right) + \beta_7 \left(intrat_t\right) \\ & + quarter_2 + quarter_3 + quarter_4 + u_t, \end{aligned} \tag{4.10}$$

where t = 2000Q1, 2000Q2,.....,2018Q4 determines a quarters in different years and u_t are unobserved variables.

First difference in TS

If data are highly persistent, it is also appropriate to use the first difference in TS. First differences are an alternative to the use of time trend or the OLS regressions. In this case, I state that the OLS has a unit root (see Table 4.3). Thus, they are highly persistent. The first difference also fixes the serial correlation in many cases and (Wooldridge 2012). I do not test for the serial correlation of the first difference regression again. However, I apply the HAC standard errors as I do for all regressions.

I will use this model in further analysis with the following form:

$$\begin{aligned} apartp_t - apartp_{t-1} = &\beta_1 (finapart_t - finapart_{t-1}) + \beta_2 (strapart_t - strapart_{t-1}) \\ &+ \beta_3 (avgwage_t - avgwage_{t-1}) + \beta_4 (gdpcap_t - gdpcap_{t-1}) \\ &+ \beta_5 (unemp_t - unemp_{t-1}) + \beta_6 (bplotp_t - bplotp_{t-1}) \\ &+ \beta_7 (intrat_t - intrat_{t-1}) + (u_t - u_{t-1}), \end{aligned} \tag{4.11}$$

where t = 2000Q1, 2000Q2,.....,2018Q4 determines a quarters in different years and u_t are unobserved variables.

4.3 Penalised Least Squares approach

Regarding the discussion about the HAC robust standard errors, I would like to list reasons, why to apply another method on these regressions. Wooldridge (2012) says that if there is a substantial serial correlation, and the sample size is small, then HAC robust errors do not have suitable properties for time series. Moreover, correcting the standard errors sometimes change the coefficients significance. In order to find a better way, how to express the best predictors, I will introduce penalised least squares estimators, particularly then least absolute shrinkage and selection operator (LASSO).

The penalised least squared (PLS) methods are useful in the case when we have many explanatory variables and less data (see Table 3.2 and Table 3.3).[xvi]Ridge, lasso, and elastic net regressions belong to the PLS. The PLS regressions choose the most important variables from the causal relationships and simplify the model by variable selection (Nasekin 2013). causal relationships

Variance Inflation Factor

Using the PLS estimators would be useful in our data since the variance inflation factor (VIF) suggests strong collinearity for almost all variables (see Table 4.5 and Table 4.6). Thus, we do probably need to have less explanatory variables in our models since we would like to avoid over-fitting. The VIF is examined by using the R_j^2. In other words, R_j^2 is the R^2 which we would obtain if we regressed x_j against all other independent variables. Sen and Srivastava (1990) define VIF as

$$VIF_j = \frac{1}{1 - R_j^2} \tag{4.12}$$

Obviously from the definition of R^2 and these two equations above, values close to one indicates "no collinearity" and on the contrary, a higher value means collinearity of the particular variable with other explanatory factors (Sen and Srivastava 1990). Table 4.5 for panel data, and Table 4.6 for time series, respectively show the values of VIF for explanatory variables.

[xvi]Panel data: n=224-252, x=11 ;Time series: n=76, x=7

Academics do not decidedly say when the variable is considered to be noted as "multicollinear". However, many sources agree that the variable is considered to be collinear with R^2 equal to 80% and VIF equal to 5, which is valid in most cases for both, panel data and time series variables.

Table 4.5: The Variance Inflation Factor for Panel data

FDI	Interest rate	Avg. wage	Unemployment	GDP capita	Population density
28.831	3.252	23.722	2.703	63.965	30.659

Finished apart	Started apart	Divorce rate	Marriage rate	Natural growth	Net migration
10.398	7.709	18.778	20.895	2.909	3.877

Table 4.6: The Variance Inflation Factor for Time Series

Interest rate	Average wage	Unemployment	GDP capita
5.046	39.012	7.214	11.320

Building plot price	Started apartments	Finished apartments
22.649	1.902	2.151

LASSO regression

In this thesis, I will use only lasso regression, firstly introduced by Tibshirani (1996), since lasso shrinks coefficients and conducts the variable selection as well. On the other hand, ridge and elastic net regression only shrink coefficient, which is not the aim of this research. I want to select the essential determinants which have the most substantial effects and which would make the interpretation of results easier (similarly as Ioannidis et al. 2018).

Moreover, least-squares estimates tend to have low bias but large variance. It is important to find a correct trade-off between "bias" and "variance" error (Fortmann-Roe 2012a). By shrinking some coefficients to zero, I expect a decrease in the variance of predicted values and avoiding possible over-fitting, which is caused by many explanatory variables. The over-fitting worsen predictive accuracy. Thus, the goal is to reduce the number of independent variables (Hastie, Tibshirani, and Friedman 2008).

Tibshirani (1996) defines the lasso estimate as:

$$\widehat{\beta}^{lasso} = arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - B_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2, \ subject\ to \ \sum_{j=1}^{k} |\beta_j| \leq \lambda, \tag{4.13}$$

where $\lambda \geq 0$ is a parameter that controls the amount of shrinkage. If λ is increased, then we obtain a greater amount of shrinkage. In the case of $\lambda = 0$, we get a standard OLS estimator. On the other hand, the number of zero coefficients increases when λ reaches higher values.

According to Equation 4.13, it is essential to choose the best value of λ. I choose λ by using glmnet package in R, which is also used in (Friedman, Hastie, and Tibshirani 2010).[xvii] The aim of the lasso is to reduce the standard errors, which are often higher if we use the HAC robust standard errors. Therefore, we want to choose *lambda.min*, as the inflection point of the lowest RMSE among the hundred lambdas estimated by cross-validation (see Figure 6.3, Figure 6.5, Figure 6.7, Figure 6.9, Figure 6.11, and Figure 6.13 in section). The value of *lambda.min* is at the left vertical dashed line, which is determined after cross-validation of our sample. The horizontal button line shows *log(lambda)*, whereas the horizontal top line determines the number of explanatory variables at the particular level of *lambda*.

However, I decided to use *lambda.1se*, which is shown as the right vertical dashed line in Figure 6.3, Figure 6.5, Figure 6.7, Figure 6.9, Figure 6.11, and Figure 6.13. *Lambda.1se* is *lambda.min* within one standard error from the minimum (Friedman, Hastie, and Tibshirani 2010). For the purpose of this thesis, *lambda.1se* is a better choice since we obtain more zero coefficients (*lambda.1se* is higher) in the comparable value of the minimum RMSE. Thus, the results will be smoother for the interpretation due to the lower amount of explanatory variables, and the goal is to get the best predictors.

The other issue with obtaining *lambda.1se* is the cross-validation which gives us more than one result after every re-sampling. Fortmann-Roe (2012b) suggests using the cross-validation for accuracy testing, which is, in this case, choosing the appropriate value of lambda. In order to get as accurate *lambda.1se* as possible and which is not "random," I run the 5-fold cross-validation 30-times for each regression. The resulting *lambda.1se* is calculated as the average of all 30 previous results.

Best Predictors Selection

After obtaining *lambda.1se*, I will show all these lasso estimators for the models in section 5. Furthermore, I will divide time series and panel data set into train and test sets (out-of-sample) in order to choose the best models (similarly as Chan-Lau 2017; Smeekes and Wijler 2018). The train set of panel data contains all years besides the year 2017, which creates a test set. I test how the predictive power of this train set is accurate when applied to the test set. Similarly, time series data are divided as a train set of the years 2000-2016 and test set containing the years 2017 and 2018 (see some analogous in and out sample dividing in Case et al. 2004).

[xvii] More detail information about glmnet package, lambda.min, lambda.1se, and cross-validation available at https://web.stanford.edu: Glmnet Alpha.

I could use the cross-validation again for testing the accuracy of forecasts. However, the aim is to test for the last years. It tells us more about the suitability of the training sample and its predictions for future apartment prices. In this case, the training sample creates the regression based on the used data.

Root mean squares error (RMSE) is measured for all models, their train ("training error"), and test sets ("test error"), in order to determine the forecast accuracy. The same procedure used in their papers Chan-Lau (2017), Smeekes and Wijler (2018), and Xin and Khalid (2018). The RMSE is the square root of the variance of the residuals. In other words, RMSE indicates the absolute fit of the model to the data, which is a difference between actual and predicted values. Wooldridge (2012) defines RMSE in Equation 4.14:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}, \tag{4.14}$$

where y_i is i-th observed variable and $\widehat{y}_i$ is the predicted value of i-th observed variable. RMSE results are in the same units as apartment prices (CZK) and I summarise and comment in Table 5.4 (Wooldridge 2012).[xviii] The next section interprets not only RMSE, but also all the regression results.

[xviii] See more model fit statistics in this summary article available at www.theanalysisfactor.com: Assesing the Fit of Regression Models.

5 Empirical Results

In section 5, I comment on the results of regressions stated in section 4. Firstly, I compare the results of fixed effects (see for details Equation 4.4), and first differences (Equation 4.5). Moreover, I comment on the lasso penalisation and its contribution. Did the lasso meet the expectation from subsection 4.3? Based on all aspects of the discussion and the RMSE summary table, which shows the accuracy of these models, I will choose the best predictive model for panel data (similarly as Chan-Lau 2017; Smeekes and Wijler 2018).

Secondly, I compare four time-series regressions: OLS (Equation 4.8), Detrending regression (Equation 4.9) OLS with seasonal dummy variables (Equation 4.10), and last but not least first differences for TS (Equation 4.11). Then I evaluate the time series results according to a similar process as in panel data. Lastly, I examine the accuracy of the best predictors for TS and PD based on plotted results.

Besides all, I could not avoid the comparison with the other authors, especially with Hlaváček and Komárek (2009b) since it is one of the most relevant literature for apartment determinants in the Czech Republic. However, we have to look at a distance in this comparison. As mentioned, Hlaváček and Komárek (2009b) used slightly different data (independent variables and period) from more sources.

5.1 Panel Data Approach

Results for panel data are summarised in Table 5.1. If we focus on R-squared as a measure of fit, we can see that the first difference regression is poorly explained compared to the fixed effects. However, it is misleading to compare the R-squared of these two models since we have different dependent variables in this case. The total sum of squares (denominator for the calculation of R-squared) of the first difference in apartment price is likely to be smaller than the fixed effect of the apartment price.[xix] Thus, I will avoid making any conclusions based on R-squares (see more in Harvey 1980).

Let us focus on the overall significance of the regressions. F-statistic suggests that both models fit good enough where the first difference and fixed effect both contain seven statistically significant variables at a 10% level of significance. From this point of view, both models look similar. Nevertheless, they have some distinguish results that I will discuss later.

[xix] The absolute value of current apartment price minus average apartment price is usually greater than the absolute value of current apartment price minus apartment price of the previous period.

We cannot reject the null hypothesis (stated in subsection 3.2) for any variables that are not statistically significant in both models. Moreover, all insignificant variables were shrunk by lasso regression to zero value. Thus, I do not report them. These insignificant variables are *FDI, population density, started apartments*, and *unemployment rate*.

Explanatory Variables Results

The *divorce rate, marriage rate, population growth, average wage, and GDP capita* are statistically significant in both models and have the expected sign (see the summary of hypotheses in Table 3.1). Moreover, *GDP capita* has almost the same positive effect in FD and FE (the same result as in Égert and Dubravko 2008). Lasso regression also did not shrink the *GDP capita* coefficients to zero in both cases. On the contrary, FE suggests a stronger effect of the other four variables than FD.

Population growth is supported by lasso regressions as well in both cases and strongly increases the housing demand in the FE model. Thus, this FE result says that comparing to the previous period, every additional newborn child increases the apartment price per metre squared by more than 1 CZK. Hlaváček and Komárek (2009b) received a high significance of the *population growth* coefficient in the FE for their data between the years 1998 and 2009. They also claim that the *divorce rate* was significant as a price determinant as I do.

I can reject the null hypothesis that the *divorce rate* coefficient has no effect in favour with my expectations. The split of households significantly increases the *apartment price*. However, the lasso shrunk *divorce rate* coefficients to zero in both models. The *marriage rate* coefficient is also equal to zero after the lasso penalisation. Nevertheless, the FE and FD regressions support the hypothesis that married couples move to one apartment. Thus, they decrease the demand since they are likely to move from their apartments.

FE and FD regressions suggest that the growth of the *average wage* has a substantial positive effect on apartment prices as I assumed (see the same results in Hlaváček and Komárek 2009a). The time-demeaning lasso regression even shows that this positive effect is almost in direct proportion with the dependent variable. On the other hand, $Lasso_{FD}$ reduced the positive impact of the *average wage* on the *apartment price*. However, the *average wage* is by value the essential coefficient in the first difference lasso regression.

Table 5.1: Panel Data and Lasso Regressions: Summary Tables

	Dependent variable: Apartment price			
	First differences	Lasso_{FD}	Fixed effects	Lasso_{FE}
FDI	−0.001	x	−0.003	x
	(0.010)		(0.005)	
Interest rate	−344.146**	x	415.451***	x
	(137.000)		(158.922)	
Divorce rate	1.125***	x	1.812***	x
	(0.358)		(0.463)	
Marriage rate	−0.631*	x	−1.538***	x
	(0.377)		(0.386)	
Population growth	0.619***	0.117	1.051***	1.099
	(0.173)		(0.251)	
Net migration	0.165**	0.061	0.082	0.060
	(0.062)		(0.064)	
Population density	−16.180	x	−14.466	x
	(24.137)		(11.191))	
Finished apartments	0.322	x	0.929***	0.818
	(0.249)		(0.254)	
Started apartments	−0.532	x	−0.171	x
	(0.424)		(0.365)	
Average wage	0.636**	0.472	0.729***	0.954
	(0.277)		(0.225)	
Unemployment rate	−110.837	x	89.198	x
	(74.802)		(68.231)	
GDP capita	0.031***	0.021	0.029***	0.003
	(0.010)		(0.007)	
Observations	210		224	
R^2	0.497		0.833	
Adjusted R^2	0.467		0.823	
Residual Std. Error	1,217.777 (df = 198)		1,244.373 (df = 212)	
F Statistic	16.305*** (df = 12)		88.077*** (df = 12)	

Note: *p<0.1; **p<0.05; ***p<0.01 ; x - shrunk coefficients to zero

The most controversial variable for this model comparison is the *interest rate*, followed by *finished apartments*. The *interest rate* is statistically significant in both models with the opposite sign (-344 FD, 451 FE). Deciding on whether there should be positive or negative effect, see for example the results of the expected positive *interest rate* impact in Égert and Dubravko (2008). You can also see the apartment price growth and interest rate change in Figure 6.15, where is also suggested the negative correlation between these two variables. However, I claim that there could be a delayed effect (see the discussion in subsection 3.5). Thus, both models could have a realistic result.

The illogical result gives us a positive sign of the *finished apartments* coefficient in the fixed effects regression. Moreover, this result is supported by $Lasso_{FE}$. I stated that a higher supply of apartments should more satisfy demand. One possible explanation could be that developers tend to finish more apartments in the periods of higher housing market prices. *Finished apartments* coefficient is also positive in FD, but statistically insignificant.

The last explanatory variable that I present in this subsection is *net migration*, which is significant in FD regression, only. However, the lasso penalisation kept *net migration* in both models with the same coefficient of 0.06. Lasso models suggest that thousand more incomers comparing to the previous period, ceteris paribus, increases the apartment price per metre squared approximately by 600 CZK.

Best Performing Model for Panel Data

Deciding on a better predictive model, we have to look at Table 5.4, where is the RMSE for all regressions. The FD model has a slightly better RMSE for both the least-squares and lasso regression, respectively. The RMSE is lower also for a train set. The only part where the FE appears to be a better model is in the RMSE test for the lasso regression. Low out-of-sample RMSE is the most important for a right prediction (see the best predictors selection discussed in subsection 4.3).

When the regression is run for the years 2002-2016, how accurate does the model predict apartment prices for the year 2017? Based on Table 5.4, we can see that the lasso coefficient selection did not considerably improve FE and FD models. The FD predictive power is even lower after applying lambda shrinkage. The FE prediction, itself, slightly increased, and thus, I claim that the FE model is better than the FD after the lasso.

The FE regression has some significant coefficients that were supposed to have the opposite sign. However, the coefficient selection excluded *interest rate*, and I stated the hypothesis why the *finished apartments* coefficient would have a positive sign. Thus, in my view, time-demeaning lasso regression is the best for forecasting apartment prices. Moreover, the lasso penalisation shrunk some coefficient to zero and treated the over-fitting of the model. I will analyse this model more in subsection 5.3.

5.2 Time Series Approach

In this subsection, I will firstly comment on variables from the time series regressions, similarly as I did in subsection 5.1. Lasso regression results could not fit into the TS summary table. Thus, results are provided in Table 5.2 and Table 5.3, respectively. Secondly, I will attempt to evaluate the predictive power of TS regressions and choose the best one. Again, I will use RMSE from Table 5.4.

If we briefly have a look at the adjusted R-squared in Table 5.2, we can see that all models seem to be explained sufficiently. Nevertheless, we cannot evaluate time series models with the use of R-squared, only. There are variables as *average wage, building plot price*, and *GDP capita* that usually grow in time. Moreover, one can see that the intercept or constant is included in the summary table. Without the intercept, explanatory variables have more extreme coefficients comparing to these results. However, the apartment prices do not start from zero levels in the year 2000. Thus, it is appropriate to keep the intercepts in regressions, primarily when they are statistically significant. The only exception is the detrending regression since there is a significant *trend* that partly substitutes the constant.

Average Wage

The *average wage* is not as significant for time series as for panel data since the *average wage* is statistically significant at a 10% level for detrending and first difference models, only (Hlaváček and Komárek 2009b, they obtained a higher significance for the *average wage* in their model). An increase of *average wage* by one unit even raise *apartment price* by 1.4 units in the regression with the trend. The *average wage* in the first difference regression has a lower effect. The lasso regression in Table 5.4 kept the variable in these two models but shrank its value by more than half.

Building Plot Price

Building plot price is significant at 10% for all models which are a similar result as in Hlaváček and Komárek (2009b). *Building plot price* has a stronger positive effect on the OLS and Seasonality regressions. Table 5.3 confirms the importance of *building plot price* for our regressions. The coefficients in lasso regressions are slightly shrunk but still suggest a significant positive effect on *apartment price*.

This strong effect could be explained by looking at Table 3.3. There is shown that *building plot price* has approximately eleven times lower mean. Thus, these high coefficients perhaps compensate for a difference between apartment and building plot prices. Nevertheless, we have to be aware of the fact that *building plot price* is just as part of real estate as *apartment price*. They are part of the housing price index. Therefore, these prices are likely to behave similarly.

Table 5.2: Time Series Regressions: Summary Tables

	Dependent variable: Apartment price			
	OLS	Detrending	Seasonality	First difference
Average wage	0.102	1.408***	−0.071	0.799*
	(0.338)	(0.253)	(0.347)	(0.454)
Building plot price	9.016***	6.620***	10.167***	5.573*
	(2.494)	(1.583)	(2.675)	(3.054)
Started apartments	0.517**	0.242	0.393**	−0.071
	(0.200)	(0.162)	(0.183)	(0.084)
Finished apartments	0.260	0.156*	0.422**	0.273**
	(0.164)	(0.093)	(0.186)	(0.106)
Unemployment rate	−705.047***	−1,118.995***	−681.369***	−1,184.931***
	(178.766)	(190.203)	(176.945)	(202.106)
Interest rate	194.191	−373.382	207.748	−254.724
	(313.461)	(235.027)	(314.563)	(275.264)
GDP capita	−0.190***	−0.241***	−0.178**	−0.088
	(0.057)	(0.053)	(0.074)	(0.114)
Trend		152.488***		
		(21.878)		
QuarterQ2			573.891*	
			(310.006)	
QuarterQ3			3.853	
			(321.026)	
QuarterQ4			−611.376*	
			(358.564)	
Constant	17,296.850***	1,074.302	17,640.060***	−559.094***
	(5,310.568)	(4,183.555)	(5,395.285)	(201.793)
Observations	76	76	76	72
R^2	0.885	0.928	0.903	0.760
Adjusted R^2	0.873	0.920	0.888	0.733
Residual Std. Error	939.19(df = 68)	745.10(df = 67)	881.47(df = 65)	783.80 (df = 64)
F Statistic	74.48***(df = 7)	108.67***(df = 8)	60.41***(df = 10)	28.91***(df = 7)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 5.3: Lasso Regressions for Time Series Models

	Dependent variable: Apartment price			
	OLS	Detrending	Seasonality	First difference
Average wage	x	0.629	x	0.299
Building plot price	5.065	7.409	5.215	4.369
Started apartments	0.332	0.305	0.270	x
Finished apartments	0.169	0.155	0.291	0.006
Unemployment rate	−456.800	−832.527	−426.461	−761.339
Interest rate	x	−77.013	x	x
GDP capita	x	−0.156	x	x
Trend		72.871		
Quarter_1			157.240	
Quarter_2			2.924	
Quarter_3			−2.055	
Quarter_4			−601.840	
Constant	9,594.836	7,400.671	9,183.990	−162.996
Note:				x - shrunk coefficients to zero

Started and Finished apartments

Started apartments coefficient is statistically significant in the OLS and seasonality regressions and has the expected sign. It says that more started apartments implies the expected higher demand for housing in the future and thus, higher apartment prices. *Finished apartments* coefficient is significant for all models instead of OLS regression. *Finished apartments* variable has a positive effect on *apartment price* which I did not hypothesised. However, the lasso penalisation did not shrink the *finished apartments* coefficient, and this result is the same as for panel data models (see the discussion in subsection 5.1).

Unemployment Rate

Besides *Building plot price*, the *unemployment rate* is the next variable, which is statistically significant for all regressions. Moreover, the *unemployment rate* has the expected negative effect on the dependent variable, which is stronger for the regression with a trend and the first differences. It is opposite to positive *Building plot price* coefficients, which are higher for the OLS and seasonality regressions. This difference is compensated by the coefficient of the *average wage*, which is positive and significant again only for the detrending regression and first differences.

Taking into account Table 5.4, we can see similar results for the lasso, but we obtain a slightly lower influence on apartment prices. Therefore, I claim that the *unemployment rate* is one of the best predictors for time series data set where a 1% increase of unemployment means the decrease of apartment price between 426 and 832 CZK per metre squared according to the lasso regressions.

Interest Rate

In the TS analysis, the *interest rate* is not significant in the least-squares models. Nevertheless, after the lasso coefficient selection, the detrending model shows the negative sign of *interest rate* coefficient which was also suggested in FD regression in Table 5.1. However, the *interest rate* looks not too important in apartment price-determining for time series since the *interest rate* is involved only in one regression.

GDP Capita

The last common explanatory variable for all models is *GDP capita*. The first three models suggest a negative effect of *GDP capita* on *apartment prices*. Thus, I can reject the null hypothesis in disfavour with my expectation, and the result also goes in contradiction to panel data results. This result is astounding since the real GDP per capita has grown and declined similarly over time as the housing price index (see Figure 6.16).

The only explanation might be the case when the real GDP per capita grows faster than the *apartment price*. Consequently, a higher *GDP capita* would imply a proportionally "lower" *apartment price*. However, both Table 3.3 and Figure 6.16 show that the *apartment price* grew faster across time. In my point of view, this result only compensates for high positive constants in the OLS and seasonality regressions. In the detrending regression, the compensation might be for the significant positive *trend* and the high *average wage* coefficient.

Best Performing Model for Time Series

In the previous parts, I described the summary tables based on different explanatory variables. Now, I will more focus on the overall model's forecast quality. In the first view, I see the possible improvement in constant variables that are in general lower than in the least-squares models. Thus, these values are closer to the *apartment price* mean in Table 3.3, which might suggest a better prediction.

OLS Regression

The OLS regression has only four significant variables, whereas the coefficient of *GDP capita* is shrunk to zero in lasso regression and insignificant *finished apartments* remain after the lasso penalisation. Looking at Table 5.4, we can see that the RMSE increased for the whole regression and train set in $Lasso_{OLS}$. On the other hand, the RMSE of the test set is lower for the lasso regression, and there is our focus.

Detrending Regression

The detrending model has no zero coefficients in lasso regression. This suggests that the model is perhaps not over-fitted by itself. The most obvious change is in the intercept, which is much higher in the cost of lower trend influence after the lasso penalisation. According to the RMSE table, the prediction is better in $lasso_{Detrending}$, and the trend suggests the increase in *apartment price* by almost 300 CZK per year. In comparison to the other TS models, the detrending model contains *interest rate* and *GDP capita* in lasso regression. The detrending model has the lowest RMSE among all TS models. Nevertheless, the predictive power seems to be inaccurate.

Seasonality Regression

The lasso penalisation kept insignificant *started apartments* in the regression. The seasonality regression says that different quarters do have a substantial effect on *apartment price*. However, the lasso penalisation, which has a slightly more accurate forecast, shrunk the quarter binary variables. Thus, I will interpret the quarter variables from $lasso_{Seasonality}$.

$Quarter_4$ has the value -601 CZK, which I would classify as compensation to higher *GDP capita* and the *average wage* in the fourth quarter. Nevertheless, these two variables are shrunk to zero in lasso regression. Thus, I do not know how to interpret this coefficient. Besides quarter dummy variables, the seasonality model is still very similar to the OLS model. Therefore, I claim that we should not attach the weight to the impact of seasonality.

First Difference

The first difference regression has, as in panel data summary table, lower R-squared and less significant coefficients than the other models. The constant has a negative sign which might be confusing since if there is no change between two periods than the apartment price should decrease by 559 CZK (Table 5.2, or 162 CZK (Table 5.3). Looking at Table 5.4, it is evident that the lasso regression did not improve the FD forecast.

Based on the results in Table 5.2, Table 5.3, and Table 5.4, the detrending model looks the best among the others. Besides *GDP capita*, all the significant coefficients have the expected sign, and the lasso penalisation did not shrink any coefficient to zero which means that the model has perhaps good variance-bias tradeoff and the lasso method does not treat the over-fitting. The RMSE is also the lowest for this model.

However, the FD out-of-sample RMSE is substantially higher than the train RMSE, and the test set RMSE of the other models. Thus, the detrending model does not have a good forecast of *apartment price*, which is essential for this thesis. The best prediction has the first difference from the least-squares regressions. Nevertheless, the lasso application worsens this predictive accuracy. I would claim again that the lasso should not be used with this form of data for the first difference in TS and PD, neither.

Table 5.4: Root Mean-Squared Errors for Regressions

	Dependent variable: Apartment price			
	Data set	RMSE	RMSE Train	RMSE Test
OLS	TS	888.381	804.046	1,806.818
Lasso_{OLS}	TS	1,030.518	867.74	1,568.547
Derending	TS	699.592	557.837	3,521.380
$\text{Lasso}_{Detrending}$	TS	771.348	600.550	2,884.467
Seasonality	TS	815.183	730.625	1,626.337
$\text{Lasso}_{Seasonality}$	TS	953.750	849.500	1,604.946
First difference	TS	738.976	709.719	722.947
Lasso_{FD}	TS	862.210	631.907	1,656.884
Fixed effects	PD	1,210.583	1,190.268	1,762.448
Lasso_{FE}	PD	1,350.116	1,334.925	1,614.666
First difference	PD	1,182.472	1,156.376	1,634.556
Lasso_{FD}	PD	1,330.061	1,299.998	2,139.706

5.3 Best Predictors of apartment prices

Based on the above discussion, the best models for apartment price predictions are the lasso regression with the time-demeaning data which give us a forecast of different Czech regions, and the first difference regression with HAC standard errors which show us a forecast of the average apartment price across the country.

Besides the first difference regressions, the lasso method improved the model's predictions. Notably, the lasso shrunk seven out of twelve coefficients to zero in the fixed effects regression. This confirmed that panel data models had too many explanatory variables, and thus, the FE and FD in panel data were over-fitted. Even though the improvement of test RMSE was not too visible, the lasso selection chose the variables for the groups with similar patterns or shrunk the unnecessary coefficients to zero.

Figure 6.1 shows the deflection of lasso prediction for time-demeaning data from the actual apartment price, which is determined as a horizontal x ax. The fitted value is based on the train set (years 2002-2016) and forecasts the out-of-sample (the year 2017). The forecast is overvalued if the column is over the horizontal ax, and vice versa. The time-demeaning lasso prediction tends to overvalued apartment prices in almost all regions with the use of panel data.

The most accurate prediction is for Prague and Pardubicky region. On the other side, the most substantial deviation is for Ustecky region. This result is substantially inaccurate since the real apartment price per metre squared was only 5534 CZK in 2017, whereas the prediction suggests 8539 CZK. This discrepancy has been given by a decrease in the real apartment price since 2009 in Ustecky region (see Figure 1.1. Thus, the model based on a fixed effect regression cannot note this individual price decline.

Figure 6.2 shows the same deflection for the best time series regression. We can see that the fitted value is closest to the actual value in 2016 when the price growth was not as fast as in the following years. On the other hand, the first difference model does not precisely forecast the apartment price in the years 2017 (undervaluation), and 2018 (overvaluation). Nevertheless, the deviation is almost within one thousand CZK per metre squared, which I consider as a successful result.

One may say that there could be involved the interaction or quadratic terms when I finally determined the best models. However, I did not find any examples in the present literature. Moreover, I do not use binary variables. The different quarters are the only exception. The interaction term as the $quarter_1$ multiplied by the *unemployment rate* would be useful if we have some seaside regions where the housing demand is much lower, and the unemployment is higher in the winter quarter.

Nevertheless, it is not the case of Czechia, and the seasonality model is not chosen as the best model. I did not want to do an interaction term with two continuous variables which I consider beyond the content boundary of this Bachelor Thesis. Moreover, most of the interactions I attempted were insignificant or were ambiguous for the interpretation. Thus, the interaction terms would be used in further analysis.

Some variables would be lagged in the model. I dealt with many models, and the interpretation would have been prolonged. Moreover, I did not draw any inspiration from the existing literature. However, it might be interesting to use a model with lagged explanatory variables in further analysis.

6 Conclusion

In this thesis, we chose the econometric models for time series and panel data, which most precisely predict the price of second-hand apartments. We discussed their pros and cons. The best forecast for panel data is given by the lasso regression with the time demeaning data. The best forecast for time series is given by the first difference where the lasso regression only worsen the prediction of this model. Even though, the lasso regression improved the other time series models, this example says that the lasso regression is not convenient for all kind of data.

The regression with a trend variable had the best results among time series models in compliance with the stated hypotheses. However, the model with a trend had inaccurate forecasts. Searching for the best predictive model was the goal of this thesis. I tested the predictive accuracy on the last years observations. Several econometric models were introduced for the analysis as the fixed effects, first differences and OLS. In order to improve the efficiency and robustness of estimators, I used heteroskedasticity and autocorrelation consistent standard errors. Moreover, I used penalised estimators by applying lasso method.

For panel data, the best price determinants of apartment price are the natural population growth and average wage, where the average wage was also significant variable for time series. The building plot price and unemployment rate are the best predictors for time series. All these explanatory variables were statistically significant in almost all models and the lasso penalisation confirmed their importance.

According to the first difference regression in time series, the increse of unemployment rate by 1%, ceteris paribus, lower the apartment price per metre squared by 1 185 CZK. Whereas, the increase of building plot price by 1 CZK, ceteris paribus, implies the growth of apartment price by 5.5 CZK. On the other hand, the lasso regression for time demeaning panel data suggests that the increase of average wage per CZK and population growth per child is almost proportional with the growth of apartment price.

Besides all, the findings should be assessed with the number of limitations. Firstly, this bachelor thesis works with the CZSO data of the second-hand apartment prices. Thus, I found the price determinants for the older apartments, only. Secondly, the collected data might be inconsistent in terms of the locations within the country or region. The statistical discrepancies caused by the sample inconsistency lower the model accuracy. Thirdly, I could not use all explanatory variables that are likely to be significant for the apartment price as rent. Lastly, I could not use more observations caused by low frequency and short timeline.

Moreover, some results have to be seen at a distance. For instance, I applied the lasso penalisation with a particular value of lambda. If I had chosen a different lambda for some regressions, the results would have been slightly different in coefficients. Instead of the cross-validation, the literature does not usually specific how to choose the value of *lambda.min* or *lambda.1se*. However, the value of lambda was sometimes very different after each re-sampling of the data. Thus, I decided on multiple repetitions and averaged the resulting lambda values.

Furthermore, the time series detrending and seasonality regressions are primarily proceeded from the theoretical background presented in Wooldridge (2012) and do not have the proper literature support. Nevertheless, the detrending regression does not have accurate forecasts, and seasonal regression performs similarly as the classical OLS regression. From this point of view, the best performing models have commonly used methods, and these time series regressions were only attempts based on the theoretical roots.

A similar research could be done by extending these results for family houses and new apartments. The other approach could be more in-depth analysis of the region's level. Property price determinants generally note whether the real estate market is overvalued or undervalued based on the fundamental price level. Thus, further research could study the real estate price bubbles. Using the econometric models in the thesis, one may test the housing affordability in the Czech Republic or in the local level.

Bibliography

Bramley, Glen (2011). "Affordability, Poverty and Housing Need: Triangulating Measures and Standards". In: *Springer Science + Business Media B.V.* 27, pp. 133–151.

Čadil, Jan (2009). "Housing Price Bubble Analysis - Case of the Czech Republic". In: 1 (Prague Economic Papers), pp. 38–47.

Cai, Wenjie and Xinhai Lu (2015). "Housing Affordability: Beyond the Income and Price Terms, Using China as a Case Study". In: *Habitat International*, pp. 169–175.

Calomiris, Charles W., Stanley D. Longhofer, and William Miles (2008). "The Foreclosure-House Price Nexus: Lessons from the 2007-2008 Housing Turmoil". In: *NBER* 14294, p. 57.

Case, Bradford et al. (2004). "Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models". In: *The Journal of Real Estate Finance and Economics* 29.2, pp. 167–191.

Chan-Lau, Jorge A. (2017). "Lasso Regressions and Forecasting Models in Applied Stress Testing". In: *International Monetary Fund*, p. 35.

Coskun, Yener et al. (2017). "Housing Price Dynamics and Bubble Risk: The Case of Turkey". In: *Housing Studies*, p. 38.

CZSO (2018). "Ceny sledovaných druhů nemovitostí". In: *The Czech Statistical Office*, p. 10.

Égert, Balázs and Mihaljek Dubravko (2008). "Determinants of House Prices in Central and Eastern Europe". In: *Czech National Bank*, pp. 1–36.

Fallahi, Firouz and Gabriel Rodríguez (2014). "Link Between Unemployment and Crime in the US: A Markov-Switching Approach". In: *Social Science Research* 45, pp. 33–45.

Fortmann-Roe, Scott (2012a). *Understanding the Bias-Variance Tradeoff*. URL: `http://scott.fortmann-roe.com/docs/BiasVariance.html`.

— (2012b). *Accurately Measuring Model Prediction Error*. URL: `http://scott.fortmann-roe.com/docs/MeasuringError.html`.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: 33.1, p. 22.

Harvey, A.C. (1980). "On Comparing Regression Models in Levels and First Differences". In: *International Economic Review* 21.3, pp. 707–720.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2008). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. 2nd. Standford, California, p. 764.

Hausman, Jerry (1978). "Specification Tests in Econometrics". In: *Econometrica* 46.6, pp. 1251–1271.

Hellerstein, Judith K. and Melinda Sandler Morrill (2011). "Booms, Busts, and Divorce". In: *B E J Econom Anal Policy*, p. 23.

Hlaváček, Michal and Luboš Komárek (2009a). "Housing Price Bubbles and their Determinants in the Czech Republic and its Regions". In: (The Czech National Bank), p. 54.

Hlaváček, Michal and Luboš Komárek (2009b). "Property Price Determinants in the Czech Regions". In: *Financial Stability Report of the Czech National Bank* (The Czech National Bank), pp. 82–91.

— (2011). "Regional Analysis of Housing Price Bubbles and Their Determinants in the Czech Republic". In: 1st ser. 61 (Czech Journal of Economics and Finance), pp. 67–91.

Hylleberg, S et al. (1990). "Seasonal Integration and Cointegration". In: *Journal of Econometrics* 44, pp. 215–238.

Ioannidis, Konstantinos et al. (2018). "Problematic Internet Use as an Age-Related Multifaceted Problem: Evidence from a Two-Site Survey". In: *Addictive Behaviours* 81, pp. 157–166.

Kholodilin, Konstantin A. (2012). "Internet Offer Prices for Flats and Their Determinants". In: pp. 1–17.

Komárek, Luboš and Ivana Kubicová (2011). "The Classification and Identification of Asset Price Bubbles". In: *Czech Journal of Economics and Finance* 61.1, pp. 34–48.

Kostelecký, Tomáš and Jana Vobecká (2009). "Housing Affordability in Czech Regions and Demographic Behaviour - Does Housing Affordability Impact Fertility?" In: *Sociologický Časopis/Czech Sociological Review* 45.6, pp. 1191–1213.

Kraha, Amanda et al. (2012). "Interpreting Multiple Regression in the Face of Multicollinearity". In: *Frontiers in Psychology* 3, pp. 1–10.

Li, Jiahan and Weiye Chen (2014). "Forecasting Macroeconomic Time Series: LASSO-Based Approaches and their Forecast Combinations with Dynamic Factor Models". In: *International Journal of Forecasting* 30, pp. 996–1015.

Mikhed, Vyacheslav and Petr Zemčík (2007). "Testing for Bubbles in Housing Markets: A Panel Data Approach". In: *CERGE-EI*. 338th ser., pp. 1–44. ISSN: 1211-3298.

MRDCR (2018). "Selected Data on Housing 2017 (June 2018)". In: *Ministry of Regional Development of the Czech Republic*, p. 184.

Nasekin, Sergey (2013). "High-Dimensional Lasso Quantile Regression Applied to Hedge Funds' Portfolio". PhD thesis. Berlin: Humboldt-Universität zu Berlin.

Phillips, Julie and Kenneth C. Land (2012). "The Link Between Unemployment and Crime Rate Fluctuations: An Analysis at the County, State, and National Levels". In: *Social Science Research* 41, pp. 681–694.

Plašil, Miroslav and Michal Andrle (2019). "Tematický článek o finanční stabilitě: Hodnocení udržitelnosti cen rezidenčních nemovitostí". In: *Česká národní banka, sekce finanční stability*, p. 12.

Poterba, James M. (1984). "Tax Subsidies to Owner-Occupied Housing: An Asset-Market Approach". In: *The Quarterly Journal of Economics* 99.4, pp. 729–752.

Selim, Sibel (2008). "Determinants of House Prices in Turkey: A Hedonic Regression Model". In: *Dogus University Dergisi* 9.1, pp. 65–76.

Sen, Ashish and Muni Srivastava (1990). *Regression Analysis - Theory, Methods, and Applications*. Springer. ISBN: 0-387-97211-0.

Smeekes, Stephan and Etienne Wijler (2018). "Macroeconomic Forecasting Using Penalized Regression Methods". In: *International Journal of Forecasting* 34, pp. 408–430.

Stone, Michael E. (2006). "A Housing Affordability Standard for the UK". In: *Housing Studies* 21, pp. 453–476.

— (2010). "What is Housing Affordability? The Case for the Residual Income Approach". In: *Housing Policy Debate* 17.1, pp. 151–184.

Thalmann, Philippe (1999). "Identifying Households which Need Housing Assistance". In: *Urban Studies*, p. 36.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society*, pp. 267–288.

Wooldridge, Jeffrey M. (2012). *Introductory Econometrics: A Modern Approach*. 5th. Mason, OH: Cengage Learning. ISBN: 1-111-53104-8.

Xin, Seng Jia and Kamil Khalid (2018). "Modelling House Price Using Ridge Regression and Lasso Regression". In: *International Journal of Engineering & Technology* 7, pp. 498–501.

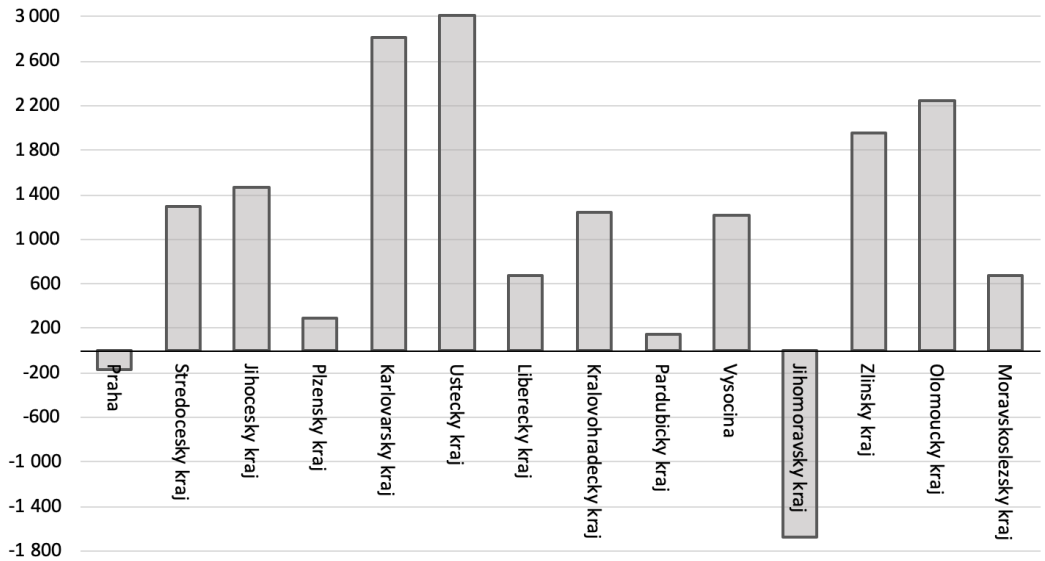
Yates, Judith and Michelle Gabriel (2006). "Housing Affordability in Australia". In: *Australian Housing and Urban Research Institute*, pp. 1–57.

Zeileis, Achim (2004). "Econometric Computing with HC and HAC Covariance Matrix Estimator". In: *Journal of Statistical Software* 11.10, p. 17.

Zemčík, Petr (2011). "Is There a Real Estate Bubble in the Czech Republic". In: 1st ser. 61 (Czech Journal of Economics and Finance), pp. 49–69.

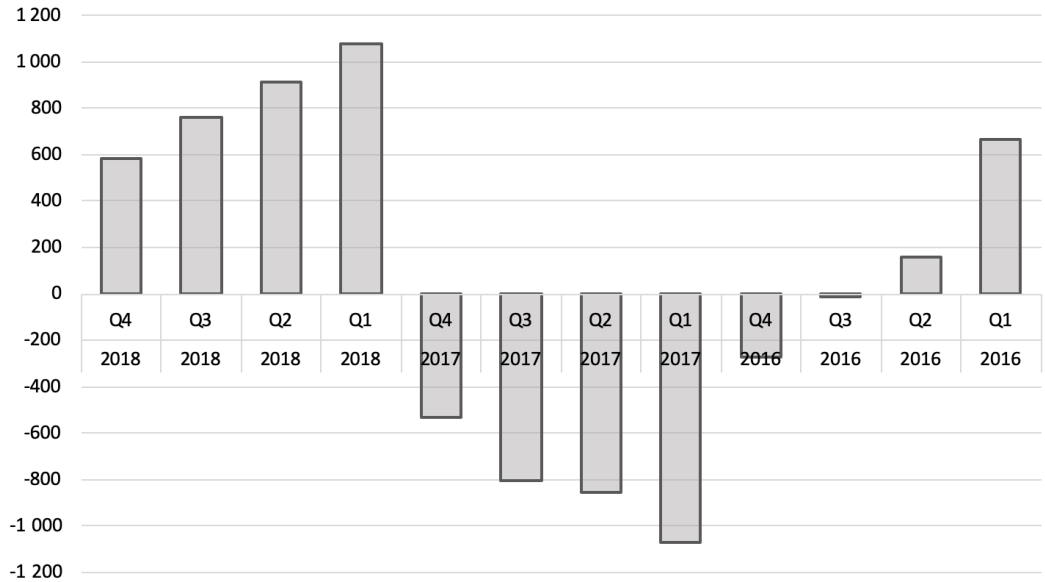
Appendix A: The Best Predictions

Figure 6.1: Deflection of Lasso prediction for Time-Demeaning Data from the Actual Apartment Price



Source: author's own analysis based on the CZSO data

Figure 6.2: Deflection of First Difference Prediction for Time Series from the Actual Apartment Price



Source: author's own analysis based on the CZSO data

Appendix B: Panel Data - First Difference

Figure 6.3: Cross-Validation for the LASSO Regression (FD Panel Data)

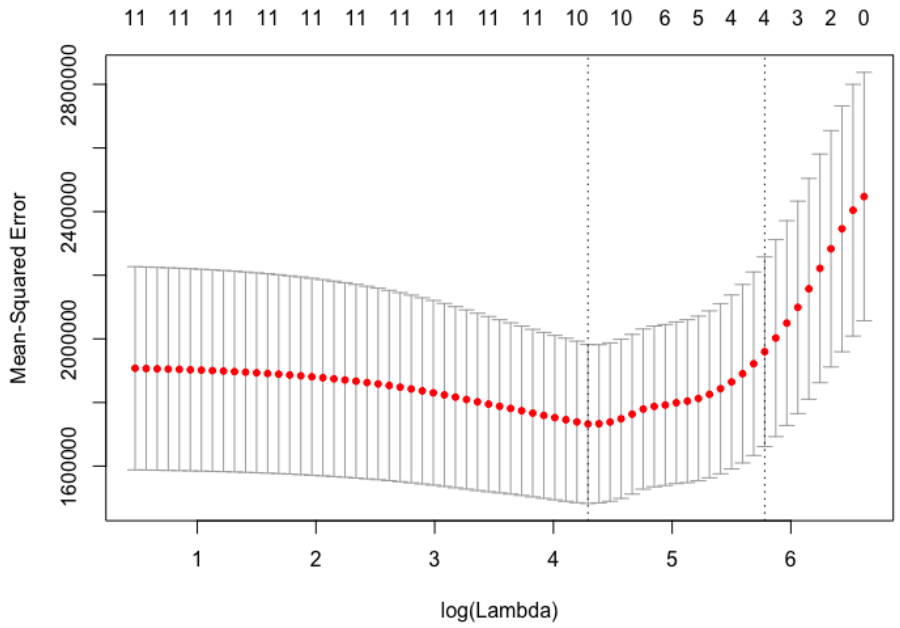
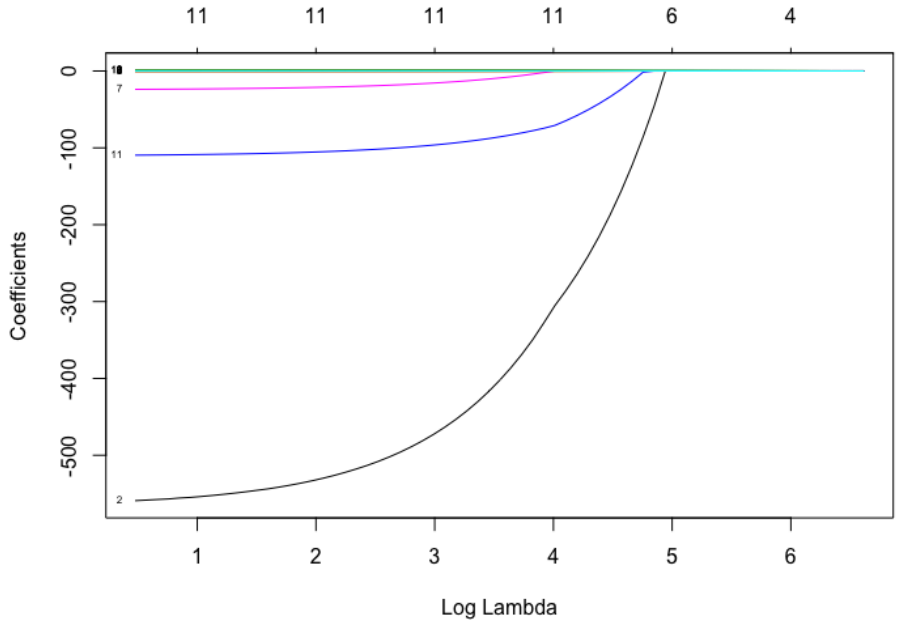


Figure 6.3: The cross-validation curves with the marked value of λ min (left vertical dashed line) and the value of λ chosen by one SE rule (right vertical dashed line).

Figure 6.4: Number of Non-Zero Coefficients Depending on the Value of lambda (FD PD)



Appendix C: Panel Data - Fixed Effects

Figure 6.5: Cross-Validation for the LASSO Regression (FE)

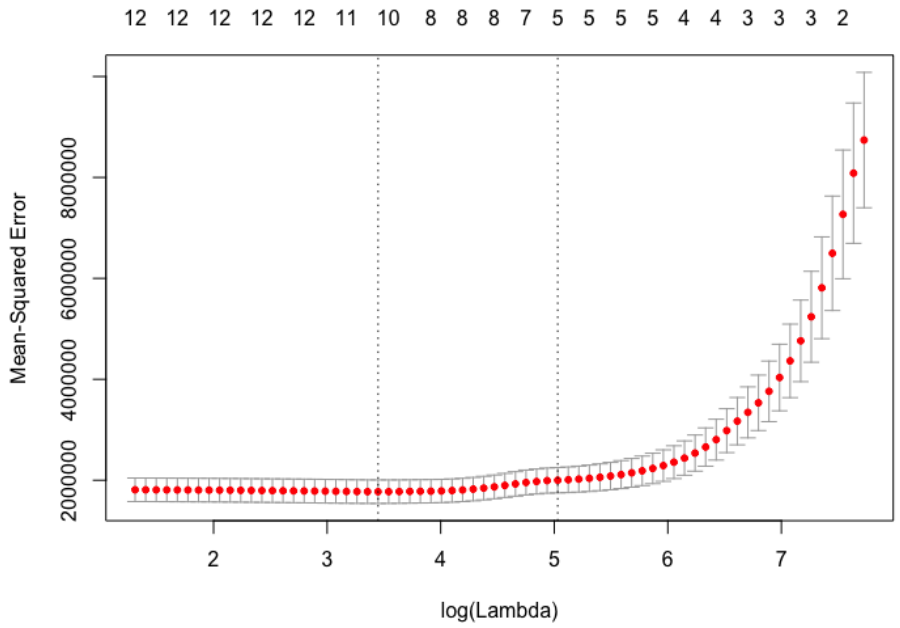
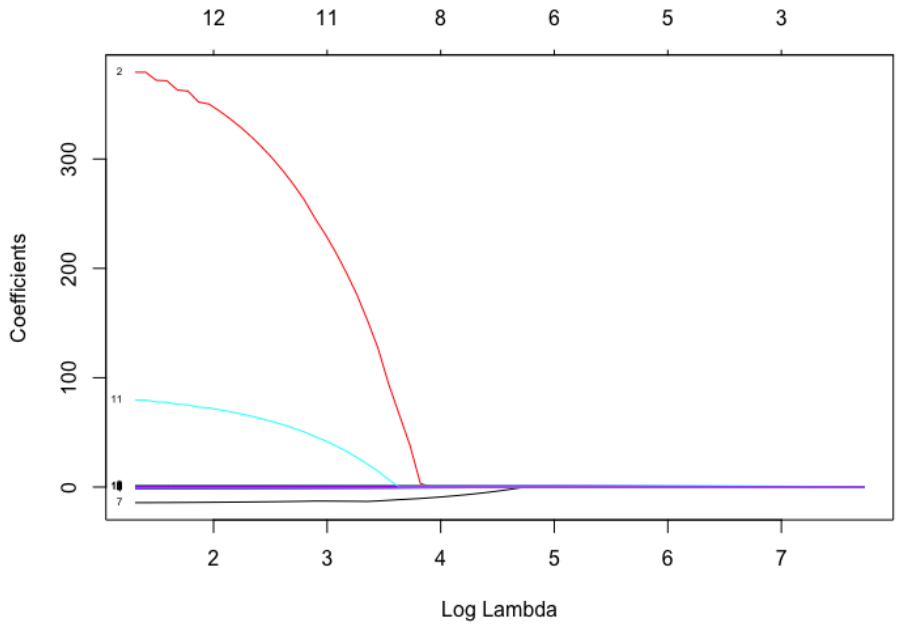


Figure 6.5: The cross-validation curves with the marked value of λ min (left vertical dashed line) and the value of λ chosen by one SE rule (right vertical dashed line).

Figure 6.6: Number of Non-Zero Coefficients Depending on the Value of lambda (FE)



Appendix D: Time Series - OLS Regression

Figure 6.7: Cross-Validation for the LASSO Regression(OLS)

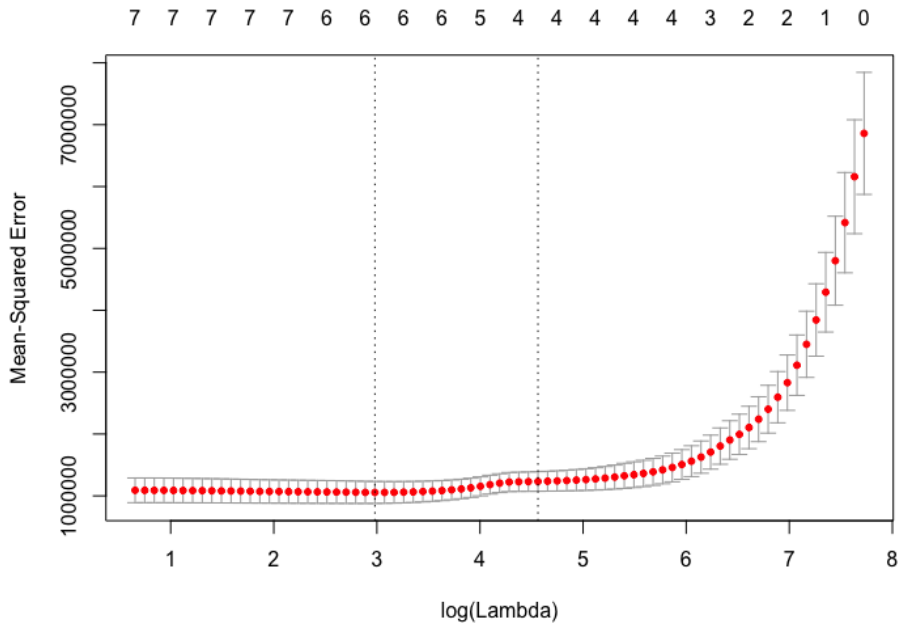
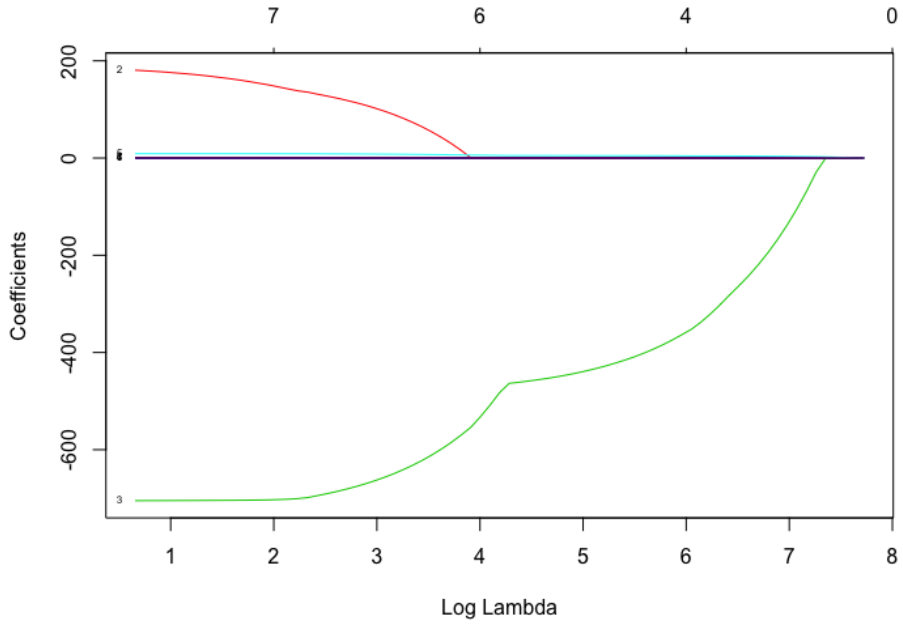


Figure 6.7: The cross-validation curves with the marked value of λ min (left vertical dashed line) and the value of λ chosen by one SE rule (right vertical dashed line).

Figure 6.8: Number of Non-Zero Coefficients Depending on the Value of lambda (OLS)



Appendix E: Time Series - Detrending Regression

Figure 6.9: Cross-Validation for the LASSO Regression(Detrending)

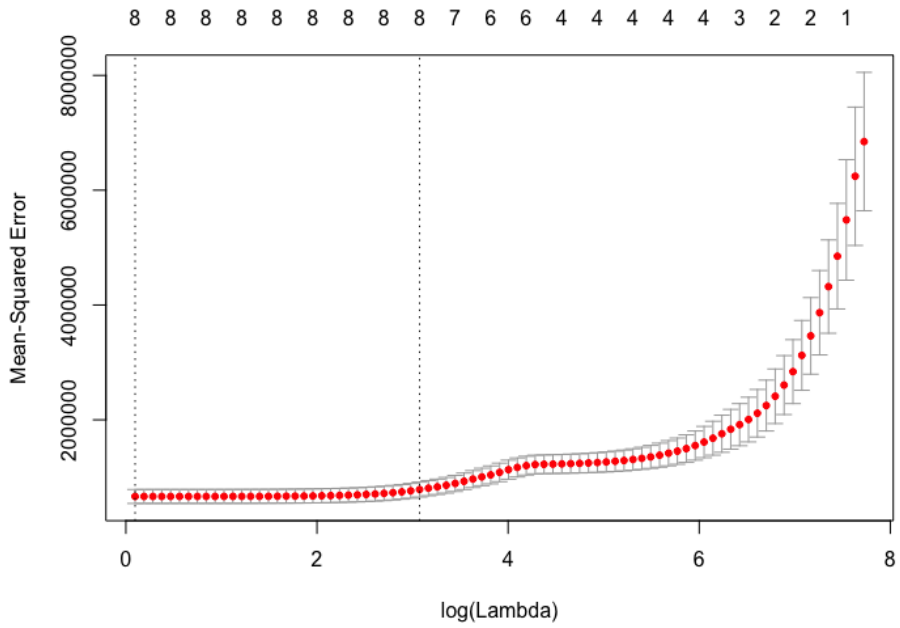
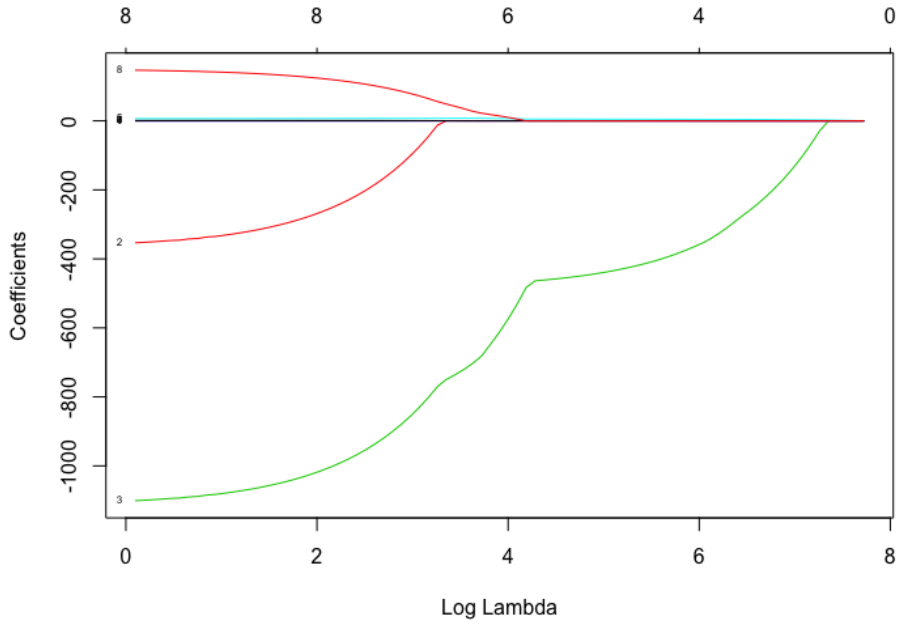


Figure 6.9: The cross-validation curves with the marked value of λ min (left vertical dashed line) and the value of λ chosen by one SE rule (right vertical dashed line).

Figure 6.10: Number of Non-Zero Coefficients Depending on the Value of lambda (Detrending)



Appendix F: Time Series - Seasonality Regression

Figure 6.11: Cross-Validation for the LASSO Regression (Seasonality)

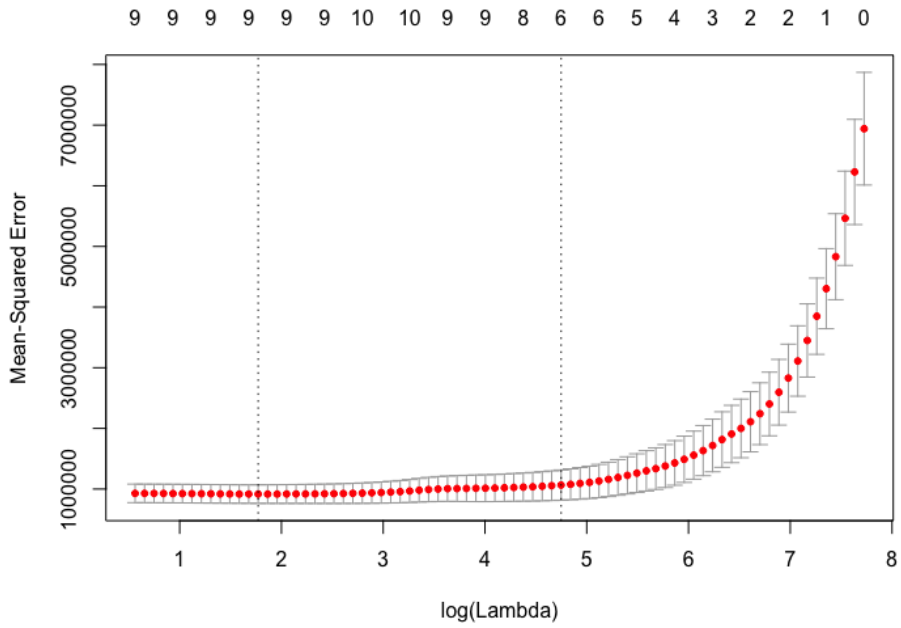
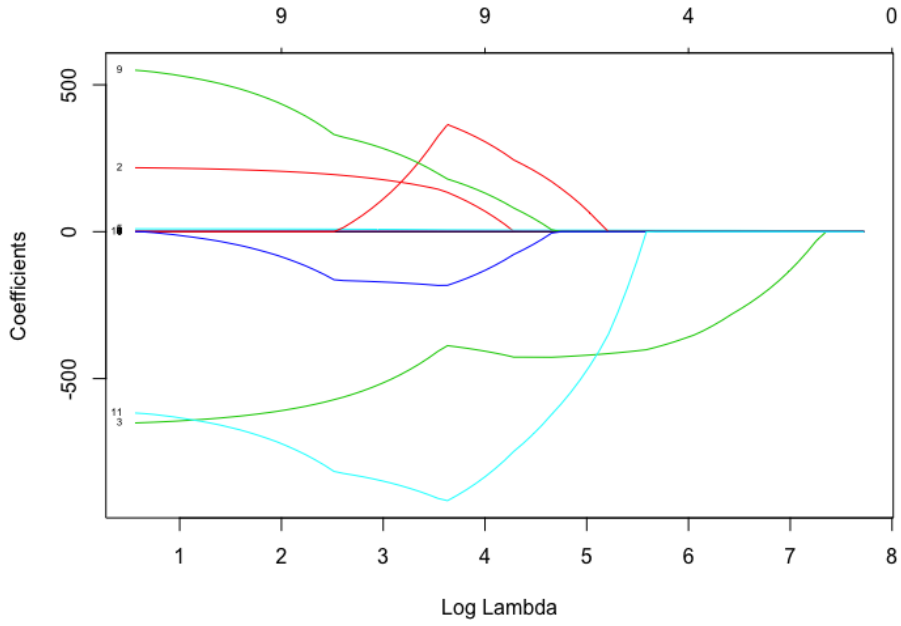


Figure 6.11: The cross-validation curves with the marked value of λ min (left vertical dashed line) and the value of λ chosen by one SE rule (right vertical dashed line).

Figure 6.12: Number of Non-Zero Coefficients Depending on the Value of lambda (Seasonality)



Appendix G: Time Series - First Difference

Figure 6.13: Cross-Validation for the LASSO Regression (FD Time Series)

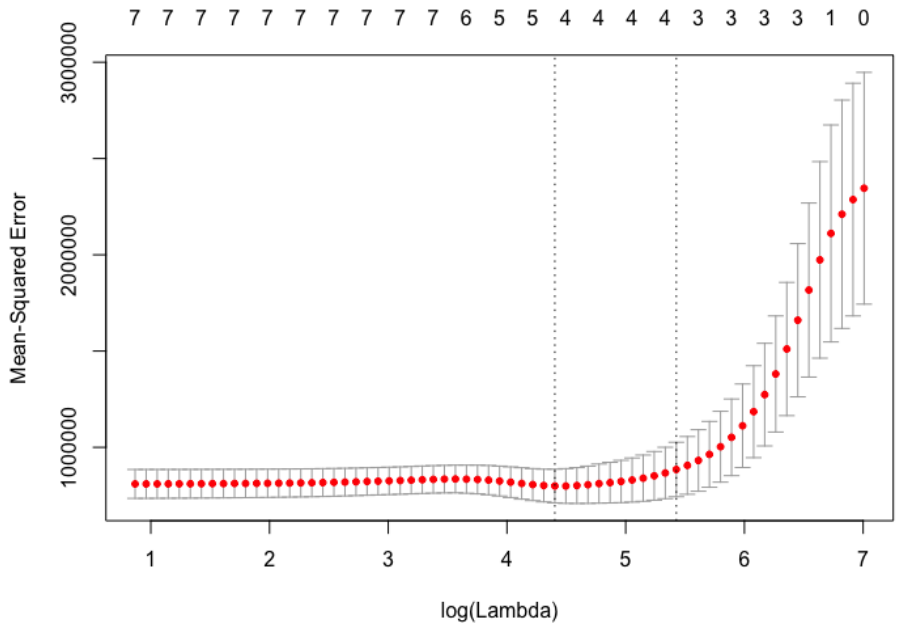
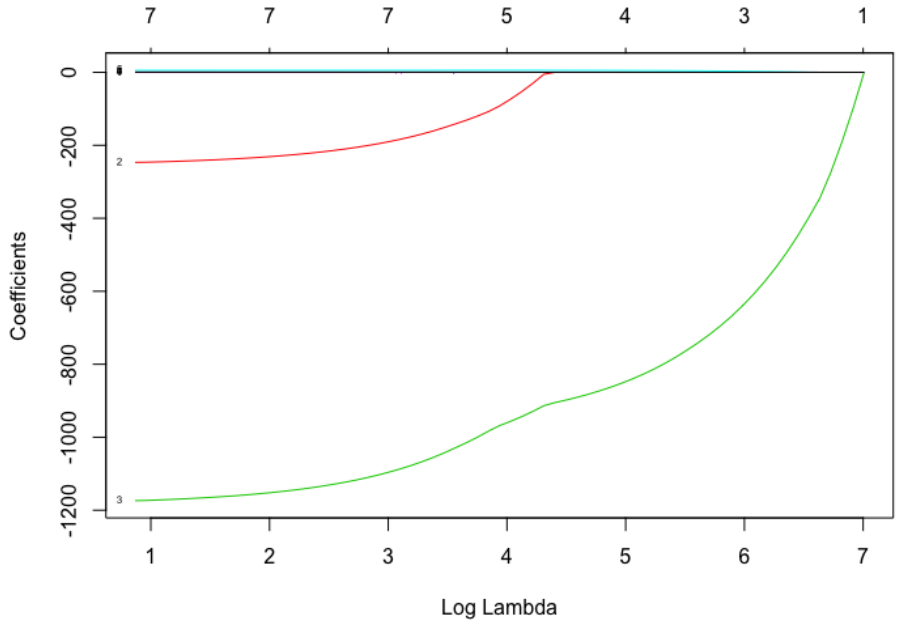


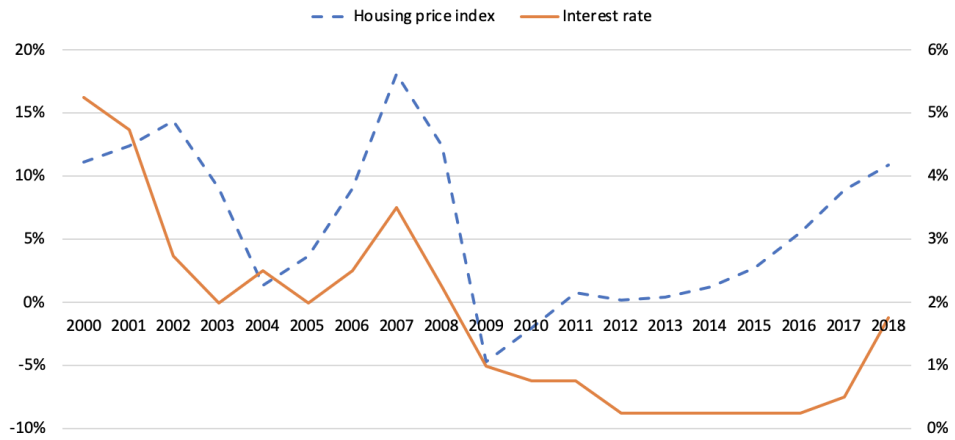
Figure 6.13: The cross-validation curves with the marked value of λ min (left vertical dashed line) and the value of λ chosen by one SE rule (right vertical dashed line).

Figure 6.14: Number of Non-Zero Coefficients Depending on the Value of lambda (FD TS)



Appendix H: Supporting Evidence

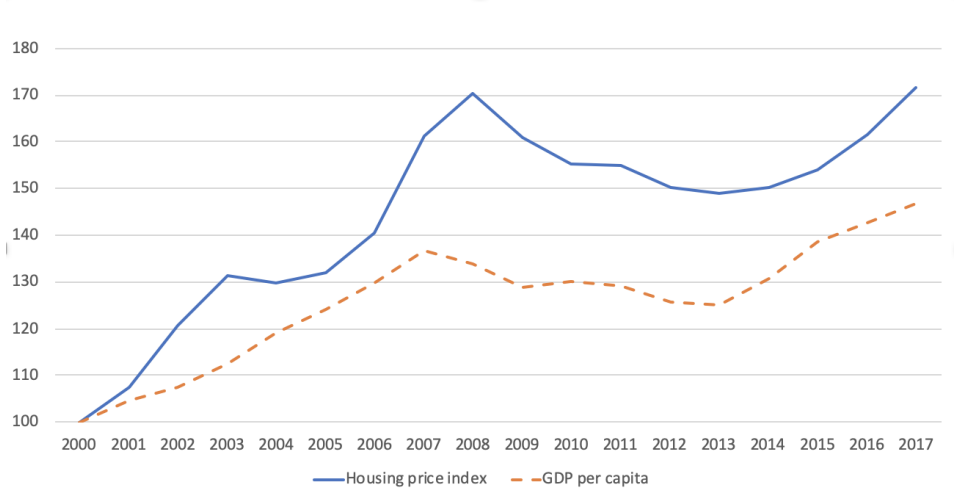
Figure 6.15: Housing Price Index Growth and Interest Rate Changes



Source: CZSO, CNB

Figure 6.15: The left vertical axe shows the percentage growth of housing price index. The right vertical axe shows the percentage change of repo rate in the Czech Republic.

Figure 6.16: Housing Price Index and GDP per Capita Growth (100=2000)



Source: CZSO