CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Bachelor thesis

2018 **Daniel Pelnář**

CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Daniel Pelnář

Price Determinants of Flats Purchased for the First Time in Prague

Bachelor thesis

Prague 2018

Author: Daniel Pelnář

Supervisor: doc. Ing. Tomáš Cahlík CSc.

Academic Year: 2017/2018

Bibliographic note

Pelnář, D. (2018). Price Determinants of Flats Purchased for the First Time in Prague, Charles University, Faculty of Social Sciences, Institute of Economic Studies. Supervisor doc. Ing. Tomáš Cahlík CSc.

Abstract

Being able to correctly estimate the true intrinsic value of a flat is important for various economic agents. This paper is concerned with the price determinants of first-time-purchased flats in Prague. It is mostly about the hedonic pricing model and its applications using data from Vivus which is one of the larger flat developers operating in Prague. Ordinary least squares was the estimation method of choice in this study. The main results are as follows. The residual analysis showed no extremely overvalued or undervalued flats based on our chosen models. Moreover, the estimated increase in prices of average sized flats in Uhříněves was 36.76% from 2017 to 2019. This is a much larger magnitude if compared with the period of the financial crisis where an average sized flat in Na Vyhlídce increased in its price by 12.83% from 2007 to 2009. It is interesting to see that even during a recession, the prices of Prague flats were raising.

Keywords

hedonic pricing model, residual analysis, OLS regression, price determinants, flat market, Prague

Range of thesis: 82 216 characters with spaces, 60 pages.

Abstrakt

Schopnost správně odhadnout skutečnou hodnotu bytu je užitečná pro mnoho ekonomických činitelů. Tento článek se zabývá determinanty cen prvotně zakoupených bytů v Praze. Převážně hovoříme o Hedonickým cenovým modelu a jeho využití kde data byla poskytnuta od jednoho z větších bytových developerů v Praze, Vivus. Metoda nejmenších čtverců byla použita. Hlavní výsledky jsou následující. Reziduální analýza nevykázala žádné nadhodnocené nebo podhodnocené byty na základě vybraných modelů. Odhadovaný nárůst cen průměrně velkých bytů v Uhříněvsi od roku 2017 do roku 2019 je 36.76%. Tato hodnota je vysoká, pokud jí porovnáme s obdobím finanční krize, kdy se cena průměrně velkého bytu u projektu Na Vyhlídce zvýšila o 12.83% od roku 2007 do 2009. Zajímavé je, že i v recesi se zvyšovaly ceny pražských bytů.

Klíčová slova

Hedonický cenový model, analýza reziduí, metoda nejmenších čtverců, cenové determinanty, trh s byty, Praha

Declaration of Authorship

1. The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

2. The author hereby declares that all the sources and literature used have been properly cited.

3. The author hereby declares that the thesis has not been used to obtain a different or the same degree.

Prague, March 30, 2018 Daniel Pelnář ____________________

Acknowledgments

I would like to express my gratitude to my supervisor doc. Ing. Tomáš Cahlík CSc. for his wisdom, experience and guidance. I would also like to thank PhDr. Radek Janhuba M.A for helping me to find errors in my Stata's code.

Last but not least, this thesis would not be possible without my family which supported me until the last minute.

Bachelor's Thesis Proposal

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Author's name and surname: Daniel Pelnář

E-mail: daniel.pelnar@gmail.com

Phone: +420 774 945 265

Supervisor's name: doc. Ing. Tomáš Cahlík

Supervisor's email: cahlik@fsv.cuni.cz

Proposed Topic:

Price Determinants of Flats Purchased for the First Time in Prague

Preliminary scope of work:

Research question and motivation

The thesis will be based on the hedonic pricing theory which states that the price of a flat is a combination of the implicit prices of the flat's characteristics. The aim of the study will be to determine the chosen characteristics and their implicit prices. Moreover, we are interested in finding out whether people are willing to pay a premium for any particular cardinal direction of windows inside a flat. Finally, the study will try to estimate the fitted values and conduct residual analysis to determine flats that are overvalued and undervalued. The findings might be important for all types of economic agents ranging from a policy maker, to an individual seeking to find the best place to live.

Contribution

Most of the Hedonic pricing literature related to Prague flat market used data from *reality.cz* or *sreality.cz* (Melichar et al., 2009; Sklenářová, 2015; Lipán, 2016), our data will come from one of the larger flat developers, Vivus (see *vivus.cz*) which operates in Prague. This will allow us to analyse the prices of flats purchased for the first time from 8 different projects in Prague. Moreover, additional explanatory variables will be added to the model and tested. For example, the cardinal direction of flats' windows or a dummy variable, basement.

Methodology

First, econometric models will be suggested based on the hedonic pricing theory. This will be followed by estimating the econometric models using ordinary least squares method. Moreover, residual analysis will be performed. There will be 10 estimated models in total because Vivus has 8 projects that are located in different locations in Prague. Furthermore, two of these projects were divided into 3 stages. Therefore we will also be estimating 2 pooled cross sections where year dummy variables, as well as, their interactions with the key explanatory variables will be present.

Outline

1. Introduction
2. Literature review
3. Data
4. Methodology
5. Results
6. Conclusion
7. References

List of academic literature:

Bibliography

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, 74. Retrieved from https://econpapers.repec.org/article/ucpjpolec/v_3a74_3ay_3a1966_3ap_3a132.htm

Malpezzi, S. (2003). Hedonic pricing models: a selective and applied review. *Housing economics and public policy*, pp. 67-89.

Melichar, J., Vojáček, O., Rieger, P., & Jedlička, K. (2009). Measuring the value of urban forest using the hedonic price approach. Regional Studies 2, pp. 13–20.

Monson, M. (2009). Valuation Using Hedonic Pricing Model. *Cornell Real Estate Review 7*, pp. 62-73

Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The Composition of Hedonic Pricing Models. *Journal of Real Estate Literature*, 13(1), pp. 1-44.

Author **Supervisor**

Contents

Chapter 1

1. Introduction

“Price is what you pay. Value is what you get.”

– Warren Buffett

Determining the true intrinsic value of a flat is important for various economic agents. It can be a potential investor seeking to allocate her capital, a concern policy maker pondering upon increasing the interest rates, or simply an individual looking for a flat to dwell in. There are many real estate valuation methods, nonetheless this thesis aims at an appraisal method based on the hedonic pricing theory.

The purpose of this study is to unveil some of the key determinants of first-time-purchased flats in Prague. Furthermore, a set of hypotheses will be tested to reveal whether, for instance, there is a premium for flats that have their windows oriented to the south. We will also attempt to quantify the magnitude of the recent increase of flat prices in Prague, and investigate whether or not some of the newly built flat units are undervalued.

Another motivation is that most of the studies that have been done on the area of the Czech Republic or Prague with regards to real estate appraisal employed data from *www.sreality.cz* which is the most frequently used real estate listing website in the Czech Republic, having more than 18 000 flat offers at the time of writing of this thesis. The website allows both real estate agencies, as well as, private advertisers to advertise on an online market place for a small daily fee. There are some possible disadvantages of these data. More specifically, measurement errors are to be expected which could cause, in some cases, the estimates to be biased. It is due to the fact that if Ordinary Least Squares (OLS) is used as the estimation method, then this is violating the zero conditional mean assumption which implies biasness of the OLS estimators (Wooldridge, 2016). This is likely to happen and it is a limitation of analysing data from *sreality.cz*. Lipán (2016) found out that some of the flat entries were exposed to human error in his study of housing prices in Prague. There are some possible ways how to reduce this measurement error, nevertheless, it is not perfect. Moreover, there was a problem with the data collection. Occasionally, it was evident that some of the flat advertisements disappeared and occurred again with an adjusted price or additional information added to the flat unit

(Lipán, 2016). This causes duplicates of observations in the data set which are not apparent and very difficult to completely deal with.

This was a good motivation for us to try to use a different source of data to analyse the flat real estate market in Prague. For this purpose, we collected data from one of the larger flat developers that is operating in Prague. We will discuss more about this in Chapter 3: Data.

This thesis is structured as follows. After this first chapter, the second chapter is presented which is concerned with the literature review of the hedonic pricing models. More preciously, the theoretical framework is developed alongside with the early studies from outside of the Czech Republic, as well as, some of the more important ones conducted with Prague real estate data. Chapter 3 talks about data collection, data source and its limitations, descriptions of data sets and variables. In the fourth chapter, the methodology is discussed, and econometric models that are going to be estimated are created. We have then a discussion about, to what extent, the MLR assumptions are satisfied, and finally an empirical framework regarding residual analysis and predictions is developed. Chapter 5 conveys the results of the estimated models. Furthermore, we try to explain the differences between the data sets and shed some light on the slightly disappointing results of the analysis of residuals. Chapter six is a conclusion with possible further research suggested.

Chapter 2

2. Hedonic pricing models in literature

In this chapter, we will carry out a literature review on Hedonic Pricing Models (HPM) focusing on real estate appraisal and flat markets. First, a theoretical foundation of HPM will be set with possible applications. Second, a brief history and studies conducted on HPM outside of the Czech Republic will be summarized. Third, some studies have already been done on HPM with data taken from the Prague housing market, we will discuss their findings in the last section.

2.1 Theoretical framework of hedonic pricing models

There are many applications of HPM, ranging from valuing cricketers in the Indian Premier League to determining whether female avatars in the fantasy game EverQuest are selling for a discount, compared to the male avatars (Karnik, 2010; Castronova, 2003). For our purposes, we will focus on the applications that are related to the real estate appraisal.

Assuming that the value of a flat is a function of all the flat's attributes, we can construct a general HPM in the following way

$$value = g(x_1, \dots, x_k) + \, v$$

where $value$ is the value of the flat, v is the error term, and $x_1, \dots, x_k$ are the attributes that determine the flat value. These attributes are usually divided into multiple categories, namely tangible flat attributes such as size or number of rooms, intangible flat attributes such as noise level or number of green spaces, and other influencing characteristics (Monson, 2009; Melichar et al., 2009).

We do not observe the value of the flat directly, therefore in most studies the transaction price or purchase price is used as a proxy, for example (Mahan et al., 2000; Melichar et al., 2009; Sirmans et al., 2005). In our study, we will also use the offer price as a proxy to the value of the flat.

Running the OLS regression on a HPM allows us to extract the estimates of flat attributes that contribute to the overall price of the flat (Rosen, 1974). There are however some limitations and assumptions that have to be dealt with in order to obtain sound results. More specifically, some of the more problematic limitations are aversion behaviour, individual perception, information asymmetry, the equilibrium assumption, and market segmentation (Vanslembrouck & Huylenbroeck, 2006).

Combining the traditional Multiple Linear Regression assumptions, together with the assumptions that address the limitations, we have the following set of the most important HPM assumptions. First, the housing market must be in equilibrium. Second, agents have perfect information and the housing market is perfectly competitive. Third, no important flat attribute is missing from the HPM, and a correct functional form is specified (Bateman, 1993; Wooldridge, 2016).

Once we obtain the estimated HPM, we can interpret the partial derivatives of each of the attributes as the marginal willingness of a buyer to pay for an extra unit of the attribute. This aids us with isolating the determinants of flat prices, allowing for the ceteris paribus interpretation (Monson, 2009).

2.2 History of hedonic pricing models

Before the Hedonic Pricing Model could have been applied for the estimation of flat prices, there had to be a paradigm shift in the standard microeconomic demand theory. This was done in 1966 by Kelvin John Lancaster when he presented his paper called, The New Approach to Consumer Theory. In his paper, Lancaster (1966) makes a distinction between consumers deriving their utilities from goods themselves – which was the former theory before 1966 – and an adjustment to this theory where he conveys that a consumer does not derive the utility from the goods themselves but rather from the bundle of attributes of which the goods are composed of (Lancaster, 1966). Hence, the consumer's problem was transformed into a similar one in which the consumer's utility is maximised over the given attributes. This is the building block that was needed for HPM.

Rosen (1974) employed the new approach of consumer theory developed by Lancaster and formally defined hedonic prices as the implicit prices of each attribute that consumers are maximising. Hedonic prices are revealed to consumers and producers from the specific quantities of each attribute and the observed prices, and can be estimated using the first-step regression analysis (Rosen, 1974).

After Rosen's 1974 paper, there is vast literature available on HPM that builds on the theories of Kevin John Lancaster and Sherwin Rosen. In the next section, we will look at some of the HPM studies that utilized housing market data from Prague.

2.3 Hedonic pricing models of housing in Prague

Recently, number of studies, related to Prague’s real estate market which employ the HPM, have been growing. Among the first ones were conducted by Jan Melichar, Ondřej Vojáček, Pavel Rieger and Karel Jedlička (2009). They studied the structural, accessibility and environmental characteristics of apartments and flats, and their impact on the prices in the city of Prague. A Czech real estate website, *reality.cz* was used as the source of their data. After obtaining 1701 observations from the period from 2005 to 2008, they developed several regression models, and discovered an inverse relationship between the flat prices and the distance from the city centre and urban forests. Moreover, the size of the flats was found to have the most explanatory power over price (Melichar et al., 2009).

Sklenářová (2015) used a different website (sreality.cz) to collect the data. This allowed for a larger set of explanatory variables to be collected and used in the HPM. The estimated models were compared with similar regressions from Israel, Spain and Indiana. The signs of the estimates were identical, but the magnitudes differed which was thought to be due to the different locations.

In order to deal with the lack of appropriate spatial inference which had not been addressed in the Prague HPM literature by then, Lipán (2016) studied the spatial models for the Prague flat market using data from sreality.cz. It was found that spatial models are more suitable for explaining the variation in prices of flats than the conventional HPM (Lipán, 2016).

Recently, there have also been some studies that were related to estimating the impact of the distance of an underground station on the flat price. Láznička (2016) investigated the effect of the metro station proximity on the apartment value for the city of Prague. The results show a positive and statistically significant impact of metro station proximity on apartment value; for an additional 100 meters away from a metro station, the apartment price decreases by 14 967 CZK (Láznička, 2016).

Our study will focus specifically on flats purchased for the first time and the determinants of their prices, utilizing data obtained from one of the larger flat developers that is operating in Prague.

Chapter 3

3. Data

We devote this chapter to explaining how data cleaning was done and we will also shed some light on the limitations and complications with the data collection.

Firstly, the source of the data will be discussed, alongside with the methods how these data were extracted, cleaned, and prepared for the analysis. And ultimately, the suitability of using such data compared to other options.

Secondly, because of the nature of the data sets, we will provide as much information as possible about each data set that is to be analysed in later chapters. This includes the number of observations and the year of approval[1] for each project.

Finally, a discussion about the obtained variables will be held – a full description of each variable includes the units of measurement, summary statistics for each project, and, in general, defining the variables so that they can be unambiguously used later in the thesis.

3.1 Data source and limitations

The data comes from Vivus (vivus.cz) which is one of the larger flat developers that operates in the capital city of the Czech Republic, Prague. This firm is registered under a company called, Pankrác, a.s[2].

The author was employed by Vivus in Summer 2016 working on the Uhříněves data set which was later used for calculating aggregate statistics and sending the information to the Czech Statistical Office (CSO)[3]. One of the benefits of this work was that the author could record any, once known, publically available information from the contracts from which the data sets were created. To be precise, it was not allowed to take owners' names, or whether or not the flat was bought by the means of mortgage, but it was permitted to record information such as the final purchase price, the number of parking lots/garages, flat's disposition, size, floor level, and so on. Most of the larger flat developers in Prague keep this information, especially the final purchase price, on their

---

[1] Or "final inspection" is an official procedure which has to take place before a flat is prepared to be transferred to the owner. From the Czech law: § 119 zákona č. 183/2006 Sb., o územním plánování a stavebním řádu

[2] See https://www.pankrac-as.cz

[3] Český Statistický Úřád, see https://www.czso.cz

websites until the flat is sold, and after that the data disappear and it is basically impossible to find them because the CSO does not require a full disclose of all these data.[4]

The other source of the data was already created data sets of the firm before the author had been employed. To the author's best knowledge, some of the data sets were exported from Vivus's website in the first few days before any flat unit was sold, and therefore contain the offer price, rather than the final price which could only be found in the contracts. This will be further discussed in the next section.

Finally, there were data sets which were written by a similar manner as the already mentioned Uhříněves, that is by going through the contracts.

We are aware that taking all the data sets from only one of the flat developers in Prague could be dangerous due to the sample selection bias (Wooldridge, 2016). A possible improvement could be made if we had an access to more than one flat developer in Prague. This is one of the possible extensions which can be done in future studies, and we will discuss it more thoroughly in Chapter 6: Conclusion.

Another possible limitation of having the data from developers is that each project is concentrated at one single location in Prague, and moreover, it is rare that there is more than one project being built at the same year. Most crucially, this limits the analysis significantly because location has to be kept constant in each project; this prevents measuring the impact of location on the flat price. Furthermore, different time periods of each project make the comparison of estimates more complicated, see Chapter 5: Results.

One possible advantage that can be expected is that the measurement error should be reduced to the minimum due to the fact that most of the data sets were used to calculate the aggregate data for the CSO, hence, they were double checked for precision. Also, there is no reason to expect any duplicates as it is the case for exporting the data from *sreality.cz*.

3.2 Description of data sets

Each data set represents one project that Vivus has administrated since 2005. Na Vyhlídce is a flat project of 3 stages from year 2007 to 2009, and Uhříněves – also a three year project with three stages. These two data sets were prepared to be analysed as pooled cross sectional data.

---

[4] Only aggregate data are required to be disclosed, see č.z 89/1995 Sb., o státní statistické službě.

Argentinská, Osadní, Luka, Aréna, Kamýk, and Pankrác are all data sets that are to be analysed by the cross sectional approach because their approval year is one year only.

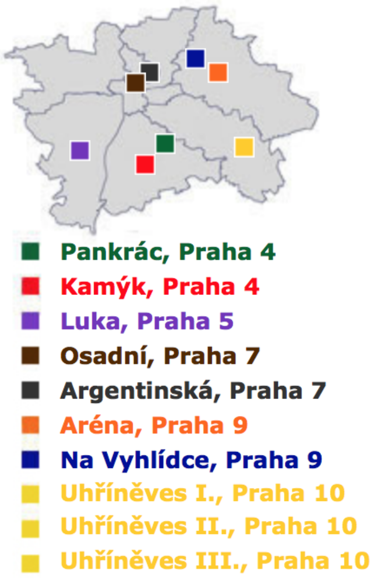
The developer could not provide private and/or sensitive information such as the name of the owner, mortgage status, the exact data of signing of SOSB[5], and so on.

Before we have a look at the data sets in more details. A table of the number of observations for each data set, and a figure depicting the locations of the projects will be presented.

Table 3.1: The number of observations

	Argentinská	Pankrác	Kamýk	Luka	Osadní	Aréna	Vyhlídce	Uhříněves
N	283	165	113	226	120	122	236	612

Figure 3.1: Locations of each project in Prague



Source: vivus.cz

---

[5] Smlouva o Smlouvě Budoucí (SOSB): a contract between the future flat owner and the developer. The cancellation penalty is between 10-15% of the flat price, depending on the project.

3.2.1 Cross sectional data sets

Project Pankrác is located in Prague 4, close to an underground station, Pankrác, and a shopping centre, Arkády. The date of approval was 28.11.2011, however, the project was selling over the next 6 years. A flat number A73-74[6] was dropped from the observations because it was a joint flat combined of two flats, and it would cause problems during the analysis. Perhaps, the most troubling thing was to adjust the prices of the flats in order to take them without taxes. For all the remaining projects, this was not a problem since they were purchased relatively fast. However, because this projects was being sold over more than 5 years, the taxes on the flat units, as well as, on the garage spots and parking lots changed for flats that were sold later. Because of some missing values for some of the SOSBs, we decided to apply the same approach as for the rest of the projects, see next section. As a consequence, the final price of some of the flats might be slightly inaccurate, nevertheless, we believe that the magnitude is negligible.

Project Kamýk is located in Prague 4 as well, nearby an underground station, Kačerov. The date of approval was 23.1.2013 when more than 90% of flats had been sold. The remaining flats were sold soon afterwards. The data set was created from the owners' contracts.

Project Luka is located in Prague 5, in a close distance from Lužina or Luka underground stations. The date of approval was 5.5.2016 with almost all the flats sold at that time. The data set was created from the owners' contracts.

Project Osadní is located in Prague 7, nearby the centre. At the time of writing of this thesis, the project is still being built, and the expected date of approval is around December, 2018. As of now, 28.2.2018, 63 flats have been sold (SOSB), which is about 51% of all flats. The cancellation fee is 15% from the price of the flat. In this project, there is no reservation option which has usually lower cancellation fee. For the sake of consistency, all prices were taken from the firm's website on the 28th of February, 2018, including the flats that already had SOSB. Hence, this data set contains offer prices rather than final purchase prices. The difference between these two is that offer prices occasionally change depending on the demand for the flats, and so this is one of the limitations of this data set that have to be taken into the account.

Project Argentinská is also located in Prague 7. This project just started to be built, and so the estimated year of approval might be the second half of 2020. As of now,

---

[6] Size = 262 m^2, price with taxes and garage spots = 22 186 117 CZK

there are no SOSB, only reservations. On the 25th of January, 2018, there were 16 (5%) flats reserved. The reservation fee is constant; not depending on the price, 57 500 CZK. Similarly to Osadní, offer prices were obtained from the website which are subject to a potential alteration.

Project Aréna is located in Prague 9, next to the O2 Aréna. The date of approval was 3.5.2013 when roughly 90% of the flats had been sold. The data set was created from the owners' contracts.

The next section talks about the two remaining projects that have 3 approval dates because they were divided into 3 stages.

3.2.2 Pooled cross sectional data sets

Project Uhříněves I.,II.,III. is situated in Prague 10. The first stage was approved on 23.1.2017. The second stage's approval date was 14.12.2017. And the third stage is still being built when the approval year is expected to be 2019. For simplicity, in the pooled cross sectional analysis, we refer to the first and second stage as years 2017 and 2018 respectively. As of now, the first two stages are completely sold and so the data were taken personally by the author from the contracts. The last stage data set contains offer prices, which complicates the analyses. We had to assume that the offer prices remain unchanged which can, but not necessary will, happen. Currently, there are 46 (23%) reserved flats in the third stage with 57 500 CZK cancellation fee. There are no SOSB. The number of observations for years 2017, 2018 and 2019 are 178, 233 and 201 respectively.

Project Na Vyhlídce I., II., III., is located in Prague 9. All three stages are sold out and their approval years were 2007, 2008 and 2009. The data set was taken from the owners' contracts and the price is the purchase price. Na Vyhlídce will be interesting to analyse because of the years when the three stages were approved. The first stage is the year prior to the financial crisis, and hence it can be intriguing to see how the estimates of prices were changing over the three years. The number of observations for years 2007, 2008 and 2009 are 94, 99 and 43 respectively.

3.3 Description of variables

In total, data were collected on 12 variables. Not all of the explanatory variables are present for each data set. Some, such as Kamýk, have only *flat_number*, *square* and *floor*. Others, like Uhříněves, have almost all the variables that are mentioned in Table 3.2. The Table depicts the availability of each variable for each project.

Table 3.2: Availability of variables

	Argentinská	Pankrác	Kamýk	Luka	Osadní	Aréna	Vyhlídce	Uhříněves
Floor	Y	Y	Y	Y	Y	Y	Y	Y
Square	Y	Y	Y	Y	Y	Y	Y	Y
Terrace	Y	Y		Y	Y			Y
Balcony	Y	Y		Y	Y			
Enclosed_bal	Y	Y			Y		Y	Y
Price	Y	Y	Y	Y	Y	Y	Y	Y
Disposition	Y	Y		Y	Y	Y	Y	Y
Orientation	Y				Y			
Basement				Y			Y	Y
Year							Y	Y
Garden				Y	Y		Y	
Flat_number	Y	Y	Y	Y	Y	Y	Y	

Note: “Y” means that the variable was collected for the particular data set. Empty cell means that the variable was not collected or the particular project did not have the attribute.

Price This variable represents either the final purchase price or the offer price, see Section 3.2. It plays an important role in the decision-making process of individuals who seek to buy a new flat. Hence it is of our primary interest. This is the explained variable in our models. It was measured in CZK without tax and additional expenses such monthly service which costs approximately 50 CZK per m^2, depending on the project.

At first, we wanted to include dummy variables for a garage spot and outdoor parking lot, however, due to personal preferences, have decided otherwise. The reason is that we do not believe that parking is one of flat’s attributes and hence it ought not to have an influence over the flat price, see Chapter 2. Due to this reason, *price* was measured without garage spots or outdoor parking. For Vivus’s projects, it holds that each flat is allocated with one parking spot. The buyer of the flat can then decide whether they will buy the parking spot or not. If there is an excess of parking spots, there is a possibility to buy more than one.

Floor *Floor* is a categorical variable, ranging from 1 up to even 10 in some projects. Each floor level is a dummy variable equal to 1 if the floor level is present, and 0 otherwise. This was done so that we could measure the effect of each floor level on price. Since all of the models will have an intercept, we would be making an error not to have a base dummy variable for *floor,* therefore first floor was set as the base year in all models. This problem is well documented and is referred to as the dummy variable trap in models with an intercept (Wooldridge, 2016).

Square A variable that measures the area of the flat in square meters. It does not include the area of a garage or an outdoor parking lot, however the area of *basement* is included. For this variable, we will also use the name *size* or *area* interchangeably.

Disposition *Disposition* is a categorical variable representing the type of the flat. Most projects have four dispositions: 1+kk, 2+kk, 3+kk, 4+kk, but there are projects that have even 5+kk and 6+kk. This is a special Czech notation which conveys the number of rooms. For example 1+kk means 1 room within which is a kitchenette. A bathroom is in a separated room. Analogously, 2+kk stands for 2 rooms within which is a kitchenette and a bathroom is in a separated room. Similarly to *floor*, 1+kk was set as the base dummy variable.

Year This is the categorical variable that was used for pooled cross sectional data sets to investigate the yearly changes in *price* by defining three dummy variables for years where the first year is the base year, and moreover interacting the year dummy variables with key explanatory variables. For Uhříněves and Na Vyhlídce, the base years are 2017 and 2007 respectively.

Orientation *Orientation* is a categorical variable indicating the cardinal directions faced by flats' windows. The basic definition follows: S=North, J=South, Z=West, V=East, JV=Southeast, and so on. In some data sets, it was complicated to define *orientation* in order not to fall into the dummy variable trap, and, at the same time, have a clear interpretation of the estimates. This is discussed thoroughly in Chapters 4 and 5.

The rest of the variables are self-explanatory, nevertheless, we will briefly define them in this paragraph. *Terrace* is a dummy variable that takes a value of 1 if the flat has a terrace, and 0 otherwise. *Balcony* is a dummy variable that takes a value of 1 if the flat has a balcony, and 0 otherwise. *Enclosed balcony* is a dummy variable that takes a value of 1 if the flat has an enclosed balcony; "lodžie", and 0 otherwise. *Basement* is a dummy variable that takes a value of 1 if the flat has a basement, and 0 otherwise. *Garden* is a dummy variable that take a value of 1 if the flat has a garden, and 0 otherwise. *Flat* *number* was added for the purpose of completeness and for being able to identify specific flat units when the analysis of residuals is performed. Using a statistical software package, Stata, a descriptive statistics table for each data set was created. All the tables can be found in Appendix A. Below, the descriptive statistics for Argentinská is presented and interpreted.

Table 3.3: Argentinská – descriptive statistics

	Mean	Median	St. Dev.	Min	Max
Price	4945922	4715000	1016568	3333000	7645000
Square	56.0976	52.31	14.39	34.23	92.22
Terrace	.0918728	0	.289358	0	1
Balcony	.6360424	1	.481989	0	1
Enclosed_bal	.2579505	0	.4382817	0	1
Floor	3.840989	4	1.848753	1	7
Disposition	1.975265	2	.6758941	1	4
Orientation	4.424028	4	2.814046	1	10

For a variable *price*, mean is larger than median and so the data are positively skewed. This holds true for a variable *square* as well. The most expensive flat is slightly above 7 500 000 CZK, and the least expensive is about 3 300 000 CZK. The flat with the biggest area is 92 square meters, and the flat with the lowest area is 34 square meters.

For the categorical variables *terrace*, *balcony*, *enclosed balcony*, the mean shows the proportion of the attribute in the whole data set. For instance, there are about 64% of flats that have at least one balcony from the whole data set.

It can also be seen that Argentinská has 7 floor levels, and should we have a look at the frequency table, see Appendix A, we will see that the most number of flats are on the second and third floor, 51 units. *Disposition* is taking four values: 1+kk (=1), 2+kk(=2), 3+kk (=3), 4+kk(=4), etc., and so it is apparent that the highest flat type is 4+kk and the lowest is 1+kk. Looking at the frequency tables, the most common flat type is 2+kk (about 61%) of all flats. Finally, *orientation* does not convey anything as of now. It has to be first properly defined. For now, we can only deduce that there are 10 combinations of cardinal directions faced by windows in each flat.

Chapter 4

4. Methodology & Empirical framework

This chapter provides an insight into the empirical framework that we are going to employ.

First, the functional form of the econometric model will be debated, and hypotheses will be stated. This will be followed with stating the MLR assumptions, and how some of them can be tested. Finally, a discussion on the residual analysis will be held.

4.1 Econometric models

There are many functional forms to choose from. Notably, the most common ones are level–level, log–log, log–level, quadratic, and interaction terms. In the following paragraph, we would like to discuss the suitability of the chosen form, regarding the log–level combination.

Malpezzi (2003) who studied hedonic pricing models prefers log forms over level–level for estimating real estate prices. According to him, there are many advantages. First, if there is a problem with heteroskedasticity, as it is usually the case with real estate data, log–level reduces this problem. Second, dummy variables can be easily interpreted in log–level model. Further, this advantage applies to all explanatory variables where the % change interpretation is convenient.

All of these characteristics are practical for our analysis of prices of new flats, especially the mitigation of heteroskedasticity, which we expect to be present; the expectation is that variance will increase with price. Moreover, the interpretation of log–level estimates is welcomed as well.

Intuitively, all the flat characteristics that could be found on the developer's website, and, consequently, in the owner's contracts, should have a bearing on the price. The reason is that the developer would not give the information about whether the flat has a balcony, or on which floor the flat is, if it was not one of the deciding factors for the potential buyer. Therefore, the general model is constructed in the following way

$$\begin{aligned}\log(price) = \ & \beta_0 + \ \beta_1 square + \beta_2 floor + \beta_3 terrace + \beta_4 balcony \\ & + \beta_5 enclosedbal + \beta_6 basement + \beta_7 garden \\ & + \beta_8 year + \beta_9 disposition + \beta_{10} orientation + v\ .\end{aligned} \tag{1}$$

In model (1), the common terminology holds; $\log(price)$ is the explained variable, and the variable of interest. $\beta_1, \ldots, \beta_{10}$ are the population parameters of the model which we seek to estimate, and β_0 is the model intercept. Then, we have the explanatory variables, and finally the error term, v.

After an estimation method is exercised – in our case, it is going to be the basic OLS – the interpretation of log–level model is as follows: If there is a one unit change of a particular explanatory variable, for example $square$, it results in a percentage increase of $100 \times \hat{\beta}_1$ for the explained variable, ceteris paribus. This means that if none of the Multiple Linear Regression (MLR) assumptions are violated, we would have Best Unbiased Estimators (BUE) and a true causal relationship between the explained and explanatory variables (Wooldridge, 2016). This is very unlikely to happen and we will talk about this more in Section 4.2.

Moreover, *floor, year* and *disposition* are categorical variables which are project specific. For instance, Argentinská has 7 floors and therefore has 6 dummy variables for each floor apart from the first one which is the base dummy variable hidden in the model intercept.

The next model (2) will help us to answer the following question. Does the size of the flat have a diminishing effect on the flat price, and is there a point of maxima?

$$\log(price) = \ \beta_0 + \ \beta_1 square + \delta_1 square^2 + \boldsymbol{x\theta} + v \tag{2}$$

The notation $\boldsymbol{x\theta}$ is shorthand for the rest of the explanatory variables described in model (1). Our primary interest will currently be $square$ and $square^2$.

Melichar et al. (2009) discovered that the size in square meters has the most explanatory power over the flat price. Moreover, it is the most commonly used variable in real estate appraisal models (Wong, Yiu, & Chau, 2013). The expectation is that the demand for larger flats is lower, which could cause the variable *square* to have a quadratic form. The model (2) will aid us with testing the following hypotheses

$$H_0: \delta_1 = 0 \,, \quad H_1: \delta_1 \neq 0 \,, \tag{3}$$

$$H_0: \delta_1 = 0 \ \land \ \beta_1 = 0 \,, \quad H_1: otherwise \,. \tag{4}$$

Expressions (3) and (4) convey how these hypotheses will be tested. Expression (3) will be done by the basic t-test, and expression (4) will be tested by joint F test of the two estimates. If the null hypothesis in expression (3) is rejected at 5% significance level or lower, then we conclude that there is a quadratic relationship between size and price. Should the null hypothesis in (3) not to be rejected, we conclude that the quadratic relationship does not exist, and we proceed to testing the null hypothesis in expression (4). If the hypothesis is rejected at 5% significance level or lower, then we keep *square* and *square*^{2} in the final model nonetheless.

Furthermore, if $\hat{\delta}_1 < 0$ & $\hat{\beta}_1 > 0$, provided that both are significant, we can infer that the relationship is a concave function and the turning point can be found, see expression (5).

$$\log\widehat{(price)} = \hat{\beta}_0 + \hat{\beta}_1 square + \hat{\delta}_1 square^2 + \boldsymbol{x}\hat{\boldsymbol{\theta}} \,,$$

$$\frac{\partial \log\widehat{(price)}}{\partial square} = 0 \,, \tag{5}$$

$$square^* = \left|\frac{\hat{\beta}_1}{2\hat{\delta}_1}\right| ,$$

the absolute sign is there because the area cannot be negative. We have denoted the turning point as $square^*$.

4.1.1 Econometric model: cardinal directions

This model focuses on the variable *orientation* and so it is more specific because we have only two data sets where this variable is present. It is complicated because the variable is defined differently in each data set. In this part of the thesis, we will create a model for Argentinská which could help us to test hypotheses related to cardinal directions.

There are four hypotheses of interest: **1)** Do people pay premium for flats oriented to the south (because these flats tend to be warm the whole day)? **2)** Do people pay premium for flats oriented to the west (because they get some sunlight after work)? **3)** Do people pay premium for flats oriented to both south and west (because they get both warm the whole day, and sunlight after work)? **4)** Which of these three have the highest premium?

$$\log(price) = \beta_0 + \beta_1 south + \beta_2 west + \beta_3 southwest + \boldsymbol{x\theta} + v \qquad (6)$$

In order to be able to test these hypotheses, model (6) was created where $x\theta$ represents the remaining explanatory variables. There are 10 combinations of cardinal directions that the windows in each flat can face. Because there are flats that are oriented to all directions, three non-intercepting sets had to be defined. The 10 dummy variables were redefined into 4 dummy variables that can be seen in model (6): $south = \{j;\ j,s;\ v,j;\ j,v,s\}$, $west = \{z;\ z,s;\ z,v\}$, $southwest = \{z,j\}$, and $\text{baseDummy} = \{v;\ v,s\}$. The frequency table for orientation, can be found in the Appendix A.

The testing procedure for the first three hypotheses is analogues to what was done in the previous section. The testing of the fourth hypothesis is a simple comparison of magnitudes of the estimated coefficients.

4.1.2 Econometric model: time interactions

The last model will deal with our pool cross sectional data sets; Uhříněves, and Na Vyhlídce. In order to be able to estimate a pooled cross sectional model, we need n-1 time dummies where n is the number of periods (Wooldridge, 2016). The basic model for Uhříněves can be expressed as follows

$$\log(price) = \beta_0 + \beta_1 y18 + \beta_2 y19 + \boldsymbol{x\theta} + v \qquad (7)$$

where $y18$, and $y19$ are year dummy variables 2018 and 2019 respectively. The base year is 2017 and is part of the intercept, β_0.

We would also like to know what happens with the size of the flat in each year. This can be easily implemented in the model by adding interaction terms, see model (8).

$$\log(price) = \beta_0 + \beta_1 y18 + \beta_2 y19 + \beta_3 y18 * square + \beta_4 y19 * square + \boldsymbol{x\theta} + v \tag{8}$$

4.2 MLR assumptions

In order to draw any meaningful results from our analysis, the cross sectional and pooled cross sectional data have to satisfy the MLR assumptions (Wooldridge, 2016). Hence, all of the assumptions will be stated and the methods of testing will be discussed. Moreover, we will have a look at, to what extent, the MLR assumptions were satisfied. At first, we wanted to present the results in Chapter 6: Empirical results as it is common, however, we decided otherwise. In our opinion, it is more suitable to have the whole section about assumptions together.

MLR.1 Linearity in Parameters This is a definition which says that we have to use a population model that is linear in parameters. In other words, the population slopes; β_1, β_2 ... , and population intercept; β_0 have to be linear.

The general model, as well as, all the additional ones, presented in the previous sections, are linear in parameters.

MLR.2 Random Sampling Obtaining a random sample that follows the population model. This assumption is problematic. The reason is that our data are from only one flat developer, hence, it is difficult to argue whether or not the sample is random. Nevertheless, we believe it is a good representation of the whole flat population of Prague because Vivus's projects are scattered all over Prague.

Apart from omitting one joint flat, all the other flats of all the projects that Vivus has ever built are in the data sets so the sample could also be viewed to follow a flat population model of Vivus.

MLR.3 No Perfect Collinearity It has to hold that none of the explanatory variables are constant. Nor are there exact linear relationships amongst the explanatory variables.

We can use the Variance Inflation Factor (VIF) to test MLR.3 formally. First, from the basic model, we regress each explanatory variable on the rest of the explanatory variables, and keep all the R_i^2. Then, the VIF for each $\hat{\beta}_i$ is equal to

$$VIF_i(\hat{\beta}_i) = \frac{1}{1 - R_i^2}$$

where the rule of thumb is that if $VIF_i > 10$, then there is a problem with multicollinearity (Kutner et al., 2005).

The model was constructed so that there were no variables that would be constant. This is the reason why, in each project, there is no variable that would represent a location. Additionally, a base group had to be excluded for each particular dummy variables as we have a model with an intercept. The VIF was within the defined boundaries for every model.

MLR.4 Zero Conditional Mean This assumption is satisfied if the error term has an expected value of zero given any values of the explanatory variables. If this is not the case, endogeneity arises, and the "causal" interpretation of $\hat{\beta}_i$ might no longer hold. The three common endogeneity causes are reverse causality, measurement error, and omitted variable bias.

In our model, we can never be completely certain about the reverse causality, however we can reduce the omitted variable bias by adding as many relevant explanatory variables as possible to the model. Moreover, we can test for a functional form misspecification, which is a type of omitted variable bias, performing the Ramsey RESET test.

Unfortunately, the null hypothesis was rejected for every data set. This means that we have misspecified models and the OLS estimators are biased to some extent. Despite trying to interact various dummy variables with *square* and *square*^{2}, as well as trying different functional forms, we were not able to solve this problem.

Finally, as we said before, we do not expect to have a problem with measurement errors, see Chapter 3: Data.

Assuming MLR.1 – MLR.4, our coefficients would be unbiased.

MLR.5 Homoskedasticity The error term has to have the same variance given any values of the explanatory variables. There are many ways how to test this assumption. We used two of the more common ones – White test, and Breusch Pagan test. Rejecting the null hypothesis means that the data are heteroskedastic.

After testing, we found out that all our data sets exhibit heteroskedasticity, and so we had to apply the robust standard errors. These are asymptotically valid due to the law of large numbers and the central limit theorem. The rule of thumb is that the number of observations of 100 and more should be sufficient (Bartoszyński, 2008). All of the data sets have more than 100 observations, therefore the robust standard errors should be valid.

Under MLR.1 – MLR.5, the OLS estimator would be the Best Linear Unbiased Estimator (BLUE).

MLR.6 Normality There are two conditions that have to be satisfied for MLR.6 to hold. First, the error term has to be independent of the explanatory variables. Second, the error term must be normally distributed with zero mean and variance σ^2, i.e. $\upsilon \sim N(0, \sigma^2)$. More specifically, the normal density function is expressed as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

where mean is equal to 0.

This assumption can be disregarded under the condition of having enough observations. Our data sets satisfy this condition. However, we also conducted the Shapiro – Wilk test for normal data. Unfortunately, the null hypothesis of normality was rejected for all the data sets. Informally, this is also evident if we have a look at the histogram plots of residuals, see Appendix B.

Under MLR.1 – MLR.6, our coefficients would be BUE.

4.3 Deriving the final models

Going from the general to the final model is not a well-defined task (Wooldridge, 2016). It happens often that dropping an insignificant variable makes another variable significant, and vice versa, i.e. a final model tends to hinge on the order in which explanatory variables were added or dropped. Due to this fact, we will follow certain steps with all the data sets so that we are consistent. Moreover, a variable *square* is always prioritized since it is of our primary interest.

First, we have the general model (1) with *square* and *square*^{2}. After running a regression using the model (1), we drop all statistically insignificant variables at 5% or lower, under the condition that there are one of these: *garden*, *basement*, *balcony*, *enclosed balcony*, *terrace*.

Second, we run the OLS regression again, however this time without the insignificant variables that were dropped in step 1. Now, if some of the variables are insignificant and belong to this category: *floor*, *disposition*, *orientation, year*, we do a joint T test. This means that if there are four floors, and one of them is insignificant, we test them jointly, and providing that the null hypothesis is rejected, we keep them in the model, otherwise they are dropped.

Third, the regression is run again, and we look whether the *square* and *square*^{2} are significant. If one of them is not, we do a F test and so on.

Finally, we re-estimate the model for the last time. The estimates we obtain are from the model we call, final. The final model estimates of all projects will be compared in Chapter 5: Results.

After each step – whenever a variable was dropped – we tested the new altered model again for omitted variable bias, as well as, for heteroskedasticity.

Furthermore, we will keep the same step order for hypothesis testing if not specified otherwise.

4.4 Residual analysis

Not only are residuals the estimates of disturbance but they can also be used as an aid in the purchase of a flat. By analysing residuals and their signs, we are able to infer whether a specific flat unit is overvalued or undervalued (Wooldridge, 2016).

It is not complicated to obtain the residuals. Firstly, we use the final model for a particular data set to estimate a relationship between the flat price and all the explanatory variables. We end up with an actual (observed) price and a predicted price for each flat. Then, the residuals are defined as

$$\hat{v}_i = y_i - \hat{y}_i \tag{9}$$

where y_i is the actual price, and $\hat{y}_i$ is the predicted price of a flat. From definition (9), it can be deduced that all the flats which have their residual negative are undervalued because the predicted price is higher than what they can be purchased for. Likewise, a flat is overvalued if its residual is positive. Nevertheless, this result is based on the attributes that are in the model. There could be variables that a potential buyer might find important such as the view from a flat. We might find this flat to be extremely undervalued based on our model, nonetheless, the flat may have a terrible view on the surroundings which makes it undesirable.

Having the dependent variable in the logarithmic form, we have to do a few more steps to obtain the prediction $\hat{y}_i$ that is needed for equation (9). We would expect that it is possible to obtain the predicted value by simply: $\hat{y}_i = e^{\widehat{\log(y_i)}}$. However, this does not work; the underestimation of the expected value of y will occur (Wooldridge, 2016).

In order to obtain the predicted values $\hat{y}_i$, we have to add an adjustment

$$\hat{y}_i = e^{\left(\frac{\hat{\sigma}^2}{2}\right)} e^{\widehat{\log(y_i)}} \,. \tag{10}$$

Under MLR.1 – MLR.6, the predicted value $\hat{y}_i$ is consistent. Nevertheless, it is problematic because we have already determined that the Normality assumption does not hold for our data. Therefore, we need a further adjustment

$$\hat{y} = \hat{\theta} e^{\widehat{\log(y)}} \,. \tag{11}$$

Now we need to estimate θ but because we cannot rely on the Normality assumption, we have to use a method of moments. Fortunately, we can use the fact that $\theta = E(e^u)$ and

$$\hat{\theta} = \frac{\sum_{i=1}^{n} e^{\hat{u}_i}}{n} \tag{12}$$

where $\hat{\theta}$ is a consistent estimator of θ (Wooldridge, 2016).

This will help with obtaining the predicted value of $\hat{y}$ without MLR.6. The final step is to obtain the predicted value $\hat{y}_i$ and after we put it into formula (9), we will arrive with a residual that we can use for determining whether the given observation is undervalued or overvalued.

Chapter 5

5. Empirical results

In total, we have eight data sets from which six are cross sectional data, and two are pooled cross sections. This chapter focuses on reporting and interpreting of our findings. First, we carry out the hypothesis testing. This is followed by a section on comparisons between all the cross sectional projects and all the pooled cross sectional projects where the final estimated models are presented. Finally, we come back to the residual analysis and predictions.

Even though, it would be logical to do the interpretation and comparison of the estimated models at the beginning of this chapter, we decided to present the results of the hypothesis testing first because we wanted to have the final estimated models based on the results of the hypothesis testing.

5.1 Hypothesis testing

The hypothesis testing was done based on Chapter 4 where the process was described. Further, we use the step order discussed in Section 4.3.

The first hypothesis of interest was whether the size of the flat has a quadratic relationship with price. This hypothesis is true for every project, apart from Aréna which has the estimates of *square* and $square^2$ insignificant, see Table 5.1. Nevertheless, these two variables were kept in the final model because they were jointly significant at 1% significance level.

We were also interested in the turning point or the point of maxima. Looking at the data sets, there is one flat in Pankác whose size is 133 square meters which is 10 square meters above the turning point. Another flat is in Na Vyhlídce which has 20 square meters more than the turning point is. From all of the data sets, the most flats that are above the point of maxima are from Luka. There are 25 of them ranging from 81 to 84 square meters. In general, this result is very interesting because it means that some of the larger flats were sold cheaper than a smaller ones if we normalize them by size, and keep other variables fixed.

Table 5.1: Estimates and points of maxima

	Argentinská	Pankrác	Kamýk	Luka	Osadní	Aréna	Vyhlídce	Uhříněves
Square	0.019***	0.037***	0.029***	0.061***	0.021***	0.005	0.024***	0.033***
	(0.001)	(0.002)	(0.002)	(0.003)	(0.002)	(0.009)	(0.002)	(0.004)
Square^2	– 0.00007***	– 0.00015***	– 0.00009***	– 0.00038***	– 0.00005***	0.00005	– 0.00006***	– 0.00015***
	(0.00001)	(0.00001)	(0.00002)	(0.00002)	(0.00001)	(0.00006)	(0.00001)	(0.00003)
Maxima	136	123	161	80	210	∅	200	110

Note: Robust standard errors in parentheses. Estimates and standard errors rounded to 3 decimal places, apart from square^2. Maxima measured in m^2. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Second set of hypotheses were about Argentinská discussed in Section 4.1.1. We asked whether there was a premium for flats oriented to south, west and southwest, and which of them was the highest. The p-values are 0.67, 0.104 and 0.73 respectively, which means that none of them are significant, nor are they jointly significant; the p-value is 0.38. One possible explanation for this result is that people do not take into consideration the cardinal direction when they are purchasing a flat. Equation (13) is the estimated model that was used for testing those hypotheses.

$$\begin{aligned} \log(\widehat{price}) = \ & 14.4 - .002\, s - .008\, w + .004\, sw \ + \boldsymbol{x}\hat{\boldsymbol{\theta}} \\ & (.422)\ \ (.004)\ \ \ (.005)\ \ \ \ (.011) \\ & n = 283, R^2 = 0.98 \end{aligned} \tag{13}$$

where robust standard errors are in parentheses. South, west and southwest are denoted by s, w and sw respectively. $x\hat{\theta}$ represents the rest of the estimates.

5.2 Estimated models: comparison and interpretation

We start with comparing the final estimated models of our cross sectional projects. Table 5.2 (page 28) shows all the explanatory variables in the first column, the name of the data sets in the first row, and estimates together with robust standard errors and significance as asterisk. The explained variable is price in logarithmic form.

In general, it is difficult to compare the final models because they were estimated by data that came from different years and different locations in Prague. Nevertheless, some comparisons are possible to draw where we can try to explain the differences in magnitudes of the estimates by pondering upon the year and location. We will not be able to quantify the estimates’ magnitudes preciously, but rather deduce the general direction of their signs.

Square Flat size, denoted as *square*, was expected to be the key explanatory variable for many reasons. The first one, which we have already mentioned, is that Melichar's et al. (2009) study on real estate appraisal conveys that this variable has the best explanatory power over price. As an experiment – for each project – we regressed price on only size to see what would happen with R squared and the significance of size. For all projects, size was highly significant and R squared was still between .75 to .95. This suggests that flat developers in Prague are pricing flats primary based on square meters.

For all the projects, with an exception of Aréna, a negative parabolic relationship can be seen, which implies a diminishing effect of size on flat price. Luka has the highest estimate where an additional square meter results in approximately 2% increase in a price of an average sized flat of 54 square meters, ceteris paribus.[7] The lowest, if we do not take into account Aréna, is Argentinská where an additional square meter causes an average sized flat of 56 square meters to increase in price by about 1.1%, ceteris paribus.[8]

It appears that Argentinská and Osadní which are the closest to Prague 1, have the lowest estimates. This suggests that people might value flats that are further away from the city center. We are not completely convinced by these findings because there are not consistent with the literature, for example (Sklenářová, 2015), and furthermore there might be other factors that are not in our models and which could potentially be correlated with size.

Some of the differences could also be contributed to the different year and location. If our assessment is correct and developers are pricing flats per square meter, then they most likely have their cost calculated per square meter too. This does not explain the location because buying a land nearby Prague 1 is more costly which implies that flat price per square meter would be more expensive. However it could shed some light on the cost of hiring a construction company that would build the project. The amount that developers pay varies depending on the year, and furthermore on the choice of the construction company. Unfortunately, we do not have this information available and so we cannot draw any meaningful inferences.

Floor The expectations about the magnitude and sign of *floor* were met in the first four projects where we see that, the higher is the floor, the more expensive the flat becomes relative to the first floor. There are some very extreme values, for instance there

---

[7] $100\% \times (0.061 - 2 \times 0.00038 \times 54)$

[8] $100\% \times (0.019 - 2 \times 0.00007 \times 56)$

is a premium of 29% for a Pankrác's flat located on the seventh floor compared to the first floor, ceteris paribus. We believe that even this value is realistic due to two reasons. First, it is a well-known fact that the probability of being burgled decreases with each additional floor level. Second, in all projects, there are very efficient elevators so the time lost is minimized.

Occasionally, some of the floor levels were not significant by themselves but turned out to be highly significant when tested jointly. Moreover, the cell is empty if the particular project does not have that floor level. This procedure holds the same for the other variables as well.

The sign of *floor* estimates are negative for flats in Osadní and Aréna which means that, in case of Osadní, flats that are on the second floor are less expensive by approximately 14% compared to the first floor, ceteris paribus. Since there are also elevators and burglaries, the only possible explanation is that floor is correlated with some other factors that we do not control for in our model. This would make our estimates biased. Unfortunately, this problem persisted even when we tried to estimate the full model.

Disposition Looking at the estimates of *disposition*, we do not see anything unexpected. All the signs are positive, and mostly it is true that a higher *disposition* increases the flat price more than a lower one. For Argentinská, if a flat is 2+kk, the price increases by about 7% compared to the 1+kk flats, ceteris paribus.

Terrace *Terrace* turned out to be significant only in two projects; Pankrác and Luka. Pankrác's estimate is exactly twice as much, and the interpretation is that having a terrace, increases the flat price by approximately 13%, ceteris paribus.

Balcony In both projects which have *balcony* significant, we see that the sign is negative. This is not what we have expected. It suggests that people prefer not to have a balcony and they are willing to pay extra money not to have one. Again, we believe that a more probable explanation is that our model does not control for all the factors that might be correlated with *balcony*.

Enclosed balcony By contrast to balcony, *enclosed balcony* has the expected sign and people are willing to pay 5.1% more for a Pankrác's flat compared to other Pankrác's flats that do not have an enclosed balcony, ceteris paribus.

R squared Finally, we would like to discuss the R squared. In all the projects, R squared is very high suggesting that the final estimated models explain the variation in their dependent variables well, and they should be suitable for our analysis. On the contrary, having a high R squared does not necessarily indicate a causal relationship, but rather a strong correlation. This can occur when there is some level of violation of the MLR assumptions, which unfortunately our data might exhibit. At last, it has to be noted that in our case we cannot use R squared to decide which of our final models is the best because each of the projects have its own data set and therefore we would not be comparing the same things, see Chapter 3: Data.

Table 5.2: Estimation Results: Cross Sectional Projects

	Argentinská	Pankrác	Kamýk	Luka	Osadní	Aréna
Square	0.019***	0.037***	0.029***	0.061***	0.021***	0.005
	(0.001)	(0.002)	(0.002)	(0.003)	(0.002)	(0.009)
Square^2	– 0.00007***	–0.00015***	–0.00009***	–0.00038***	–0.00005***	0.00005
	(0.00001)	(0.00001)	(0.00002)	(0.00002)	(0.00001)	(0.00006)
Floor 2	0.024***	0.07	0.041	0.109***	– 0.139***	0.009
	(0.007)	(0.071)	(0.048)	(0.014)	(0.015)	(0.044)
Floor 3	0.038***	0.094	0.022	0.147***	– 0.124***	– 0.032
	(0.007)	(0.071)	(0.047)	(0.021)	(0.014)	(0.037)
Floor 4	0.056***	0.111	0.045	0.206***	– 0.098***	– 0.009
	(0.008)	(0.07)	(0.047)	(0.017)	(0.017)	(0.037)
Floor 5	0.063***	0.152*	0.061	0.255***	– 0.115***	– 0.033
	(0.007)	(0.072)	(0.048)	(0.018)	(0.015)	(0.035)
Floor 6	0.075***	0.154*	0.052		– 0.109***	– 0.041
	(0.008)	(0.071)	(0.048)		(0.018)	(0.041)
Floor 7	0.098***	0.29***	0.113*		– 0.133***	– 0.076
	(0.008)	(0.077)	(0.052)		(0.016)	(0.043)
Floor 8		0.239*	0.1		– 0.156***	– 0.003
		(0.094)	(0.073)		(0.018)	(0.034)
Floor 9			0.155**			
			(0.046)			
Floor 10			0.283**			
			(0.095)			
2+kk	0.069***			0.032		0.061*
	(0.008)			(0.017)		(0.026)
3+kk	0.121***			0.012		0.108*
	(0.011)			(0.015)		(0.045)
4+kk	0.13***			0.114***		0.05
	(0.017)			(0.029)		(0.201)
Terrace		0.126**		0.058*		
		(0.041)		(0.023)		
Balcony	–0.031***			–0.095***		
	(0.007)			(0.018)		
Enclosed_bal		0.051*				
		(0.02)				
(Intercept)	14.44***	13.37***	13.38***	12.73***	14.52***	14.42***
	(0.038)	(0.107)	(0.096)	(0.076)	(0.054)	(0.364)
N	283	165	113	226	120	122
R^2	0.98	0.94	0.95	0.98	0.98	0.84

Note: Standard errors in parentheses. Estimates and standard errors rounded to 3 decimal places, apart from square^2 and the intercept. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We now turn to the estimated models of our two pooled cross sectional projects. Both of these projects consist of three pooled cross sections. Notably, they are from different years and different locations in Prague, hence the comparison between them will be difficult like it was the case in the previous analysis.

Should we have a look at Table 5.3 (page 31), we will see the estimated models' table has exactly the same format as previously. The only difference is that these estimated models have time dummy variables and their interactions with square. Moreover, the year notation differs depending on the project. Year A is the year 2008 and 2018 that represents Na Vyhlídce and Uhříněves respectively. Year B is the year 2009 and 2019 that corresponds to Na Vyhlídce and Uhříněves respectively. The base years are 2007 and 2017. Some of the partial derivatives were slightly more complicated to calculate than before, therefore a thorough explanation of how the percentages were derived can be found in Appendix C.

Na Vyhlídce Examining, for now, only the partial estimates that are not a function of another variable, we can see that in year 2008, there is an increase in flat prices of 4.1% compared to 2007. This would suggest that in the time period before the financial crisis, the flat prices were growing. Some of the possible explanations for the price increase can be inflation, a higher demand, or low supply of flats in Prague. Nevertheless, we have to conclude that there was no change in flat price from 2007 to 2008 because the estimate is insignificant. The year 2009 is highly significant and the magnitude is much greater. If we still only take a partial estimate without taking into account *square*, a flat in 2009 costs 31.3% more than in 2007. It is very interesting to see such a large increase especially during an economic downturn. There is definitely inflation that would contribute a little bit to the increase, however there must be other factors which we can only speculate about. It could have been because a lot of developers went bankrupt, due to the high leveraging, which decreased the supply of new flats and so flat price increased dramatically.

Floor has the expected sign. If a flat is on the second floor, the price of the flat increases by 10.3% on average between all the years compared to the first floor, ceteris paribus.

Disposition was kept in the model as it was jointly significant but we cannot draw any interpretation from any particular estimate because they are insignificant. Therefore, we concluded that disposition has no effect on price between all the three years on average, ceteris paribus.

If we take an average sized flat from 2008 and add an additional square meter, it results in approximately 1.524% price increase compared to 2007, ceteris paribus. However, this change only stems from the additional square meter, and not from the year because the year estimate is insignificant.

Now, we will have look at the year 2009. If we add an additional square meter to an average sized flat from 2009, we can expect price increase of 1.271% compared to 2007, ceteris paribus. In this case, we took into account the year estimate because it is significant.

Finally, an interesting result is that an average sized flat costs 12.83% more in 2009 than in 2007, ceteris paribus. It is an enormous increase, especially taking into consideration that it was during a recession.

Uhříněves The method of interpretation is mostly analogous and all the calculations can be found in Appendix C.

Perhaps, the most intriguing result is that an average-sized flat costs about 36.76% more in 2019 than in 2017, ceteris paribus. This is a large difference between the two projects. In our opinion, the sharp increase can be mostly due to the prevailing low interest rates which make mortgages less expensive, and furthermore, the low interest rates might have an impact on the stock and bond market as well. It is possible that due to the overvalued stock market and bonds with almost zero yield, one of the few still profitable investments is purchasing a flat. This could have driven the demand high which in turn could have provided an incentive for Prague developers to increase the price.

Lastly, we would like to point out that it appears that people started to value enclosed balconies. Between 2007 and 2009, they were insignificant, but now they are becoming significant. A flat with an enclosed balcony increases the flat price by 4% on average between years 2017 and 2019, in comparison to a flat that does not have an enclosed balcony, ceteris paribus.

Table 5.3: Estimation Results: Pooled Cross Sectional Projects

	Vyhlídce	Uhříněves
Square	0.024***	0.033***
	(0.002)	(0.004)
Square^2	–0.00006***	–0.00015***
	(0.00001)	(0.00003)
Floor 2	0.103***	– 0.012
	(0.015)	(0.01)
Floor 3	0.169***	0.022*
	(0.018)	(0.01)
Floor 4	0.248***	0.041***
	(0.018)	(0.01)
Floor 5		0.05***
		(0.01)
Floor 6		0.053*
		(0.022)
2+kk	0.01	0.013
	(0.025)	(0.031)
3+kk	0.038	0.009
	(0.035)	(0.039)
4+kk	0.065	– 0.021
	(0.044)	(0.046)
5+kk	– 0.091	
	(0.073)	
6+kk	0.09	
	(0.115)	
Terrace		0.059***
		(0.013)
Enclosed_bal		0.04***
		(0.011)
YearA	0.041	0.101**
	(0.052)	(0.036)
YearB	0.313***	0.441***
	(0.058)	(0.036)
YearA × *Square*	– 0.00039	– 0.00002
	(0.00066)	(0.00066)
YearB × *Square*	– 0.00253***	– 0.00131*
	(0.00071)	(0.00063)
(Intercept)	13.52***	13.28***
	(0.063)	(0.126)
N	236	612
R^2	0.96	0.94

Note: Standard errors in parentheses. Estimates and standard errors rounded to 3 decimal places, apart from square^2, the intercept, and year interactions. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5.3 Analysis of residuals result

This final section deals very briefly with the residual analysis. After estimating the final modes and calculating fitted values, we found out that the difference between the actual observed values and the fitted values was minuscule. For instance, in Argentinská, even some of the largest residuals indicated that the particular flat unit was overvalued but by only about 0.5 – 1 CZK, which is completely negligible if we contrast it with an average flat price of almost 5 000 000 CZK. The same result is for negative residuals which point to undervalued units.

These findings are not completely surprising because we know that our estimated models have very high R squared and so the fitted line explains the variation in price very well, not allowing for residuals to be of high magnitudes.

Figure 5.1: Argentinská – Predicted line and actual observations

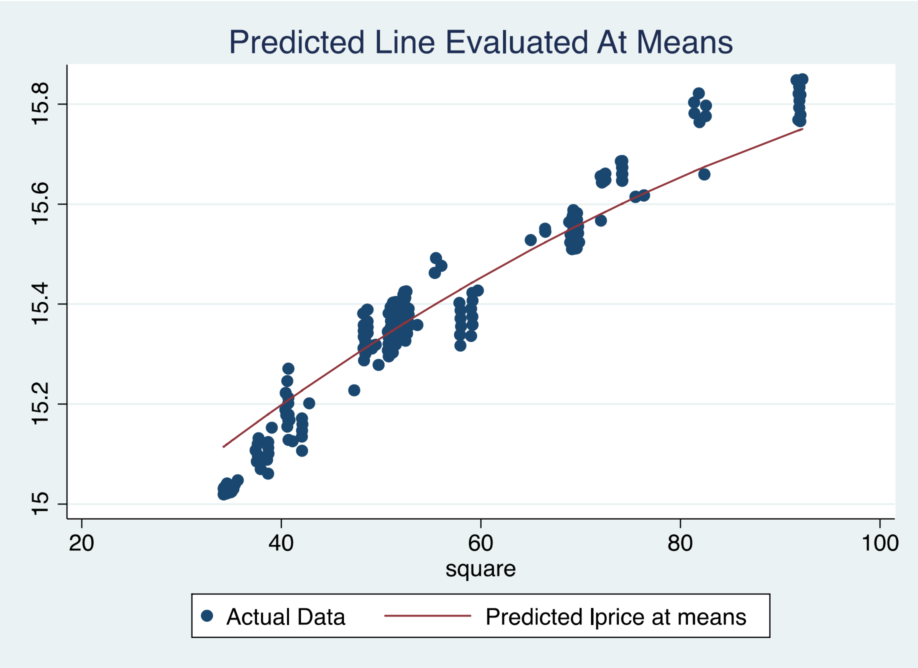


Figure 5.1 shows the actual observed values, fitted values and residuals graphically for Argentinská. Being able to create this plot in case of multiple explanatory variables, we had to evaluated the explanatory variables, apart from *square,* in terms of their means and add them to the intercept. Stata's code, as well as, the rest of the plots can be found in Appendix D.

Chapter 6

6. Conclusion

We started the thesis with a quote from one of the most famous investors of all time, Warren Buffett. He says that people ought not to make a decision based on the market or offer price alone, but rather calculate the intrinsic value of the investment.

In order to improve our understanding of the determinants of the first-time-purchased flats in Prague, we incorporated the hedonic pricing modelling. According to the hedonic pricing theory, the price of a flat is a combination of the implicit prices of the flat's attributes. After running the regressions, we tested several hypotheses and performed the analysis of residuals. One of our conclusions is that individuals most likely do not take into account the cardinal direction of the flat that they want to purchase. Another interesting result was that an average sized flat in the area of Uhříněves increased its price by 36.76% from the year 2017 to 2019, ceteris paribus where the calculations were based on the offer prices of the final year, and the purchase prices of the previous two years. Finally, the residual analysis did not show any extremely overvalued or undervalued flats based on our chosen models.

Our contributions lie in analyzing the magnitudes and signs of the flat characteristics, as well as, using different data sets which have their advantages and drawbacks. The results can be used by Prague developers to get a sense of the appropriate pricing for their future projects that are in a close proximity to one of the projects we analysed. It can also be used by a potential investor seeking to determine whether or not a given flat unit is overvalued or undervalued based on the estimates from our models. This assumes that the flat unit is from a project that is nearby one of the projects we looked into. The analysis that the potential investor can employ is called, out of sample prediction.

Notwithstanding the fact that we did not manage to find any overvalued or undervalued flats in our projects, it still does not mean that the flats are priced correctly. This was just one of many analyses that could be applied in order to estimate the true intrinsic value of a flat. Moreover, due to the possible omitted variable bias in all our models, we cannot even be completely certain how reliable the results are. In either way, we will never know.

For someone who has access, further research can be based on obtaining the data from more than one Prague developer. Alternately, it is possible to create a computer application which would scan the websites of all the larger developers. This application would be extracting the relevant data on offer prices, as well as, all the relevant attributes before the flat is sold and the data are erased from the websites. Both of these options would reduce the level of violation of MLR.2 assumption.

Bibliography

Bartoszyński, R. (2008). *Probability and statistical inference*. Hoboken, N.J. : Wiley-Interscience, A John Wiley & Sons, Inc., Publication (2nd ed.)

Bateman, I. (1993). *Evaluation of the environment: a survey of revealed preference techniques*. Centre for Social and Economic Research on the Global Environment.

Baum, C. F. (2006). *An introduction to modern econometrics using stata*. College Station : Stata Press, 2006. StataCorp LP.

Castronova, E. (2003). The Price of 'Man' and 'Woman': A Hedonic Pricing Model of Avatar Attributes in a Synthethic World (SSRN Scholarly Paper No. ID 415043). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=415043

Karnik, A. (2010). Valuing Cricketers Using Hedonic Price Models. *Journal of Sports Economics*, *11*(4), pp. 456–469.

Kutner, M., Nachtesheim, Ch., Neter, J., & Li, W. (2005). *Applied linear statistical models. Boston* : McGraw-Hill Irwin, c2005. (5th ed.)

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, *74*. Retrieved from https://econpapers.repec.org/article/ucpjpolec/v_3a74_3ay_3a1966_3ap_3a132.htm

Láznička, J. (2016). Impact of metro station proximity on apartment value in Prague. Bachelor's thesis, Charles University, Faculty of Social Sciences, Institute of Economic Studies.

Lipán, M. (2016). Spatial approaches to hedonic modelling of housing market: Prague case. Bachelor's thesis, Charles University, Faculty of Social Sciences, Institute of Economic Studies.

Mahan, B. L., Polasky, S., & Adams, R. M. (2000). Valuing Urban Wetlands: A Property Price Approach. *Land Economics*, *76*(1), pp. 100–113.

Malpezzi, S. (2003). Hedonic pricing models: a selective and applied review. Housing economics and public policy, pp. 67-89.

Melichar, J., Vojáček, O., Rieger, P., & Jedlička, K. (2009). Measuring the value of urban forest using the hedonic price approach. *Regional Studies 2*, pp. 13–20.

Monson, M. (2009). Valuation Using Hedonic Pricing Model. *Cornell Real Estate Review 7*, pp. 62-73.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), pp. 34–55.

Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The Composition of Hedonic Pricing Models. *Journal of Real Estate Literature*, *13*(1), pp. 1-44.

Sklenářová, T. (2015). Empirical Analysis of Prague Flat Market. Bachelor's thesis, Charles University, Faculty of Social Sciences, Institute of Economic Studies.

Vanslembrouck, I., & Huylenbroeck, G. V. (2006). *Landscape Amenities: Economic Assessment of Agricultural Landscapes*. Springer Science & Business Media.

Wong, S. K., Yiu, C. Y., & Chau, K. W. (2013). Trading Volume-Induced Spatial Autocorrelation in Real Estate Prices. *The Journal of Real Estate Finance and Economics*, *46*(4), pp. 596–608.

Wooldridge, J. M. (2016). *Introductory econometrics : a modern approach*. Boston : Cengage Learning, (6th ed.)

List of appendices

Appendices

Appendix A: All projects: Summary statistics

Table 3.4: Argentinská - Frequency tables:

	Frequency	Percent	Cum.
Floor:			
1.	30	10.60	10.60
2.	51	18.02	28.62
3.	51	18.02	46.64
4.	46	16.25	62.90
5.	43	15.19	78.09
6.	31	10.95	89.05
7.	31	10.95	100
Orientation:			
J	58	20.49	20.49
J,S	32	11.31	31.80
J,V,S	6	2.12	33.92
V	90	31.80	65.72
V,J	9	3.18	68.90
V,S	14	4.95	73.85
Z	26	9.19	83.04
Z,J	15	5.30	88.34
Z,S	9	3.18	91.52
Z,V	24	8.48	100

Disposition:			
1+kk	62	21.91	21.91
2+kk	172	60.78	82.69
3+kk	43	15.19	97.88
4+kk	6	2.12	100

Table 3.5: Uhříněves - Summary statistics:

	Mean	St. Dev.	Min	Max
floor	3.24183	1.544861	1	6
square	55.75593	15.99014	28.42	98.1
terrace	.1666667	.3729828	0	1
enclosed_bal	.8382353	.368536	0	1
price	2978389	925429.2	1416645	5641000
disposition	2.330065	.9550681	1	4
basement	.874183	.3319143	0	1
y17	.2908497	.4545258	0	1
y18	.380719	.4859608	0	1
y19	.3284314	.4700268	0	1

Table 3.6: Aréna - Summary statistics:

	Mean	St. Dev.	Min	Max
floor	3.54918	2.00455	1	8
square	70.30984	17.34126	52.5	145.5
price	3764070	1336779	2351950	1.01e+07
disposition	2.286885	.6861018	1	4

Table 3.7: Kamýk - Summary statistics:

	Mean	St. Dev.	Min	Max
floor	4.716814	2.350794	1	10
square	60.64363	20.05217	32.84	115.88
price	3081995	1213543	1436950	7100000

Table 3.8: Luka - Summary statistics:

	Mean	St. Dev.	Min	Max
floor	2.792035	1.345322	1	5
square	54.38673	17.14966	34.3	84.5
terrace	.4292035	.4960612	0	1
garden	.2212389	.4160024	0	1
balcony	.6150442	.487665	0	1
price	3151946	1075457	1634880	5908236
disposition	1.946903	1.164976	1	4

Table 3.9: Osadní - Summary statistics:

	Mean	St. Dev.	Min	Max
floor	4.158333	2.012548	1	8
square	66.26482	14.54506	40.56	113.08
terrace	.125	.3321056	0	1
balcony	.7916667	.4078192	0	1
enclosed_bal	.0833333	.2775443	0	1
price	6055405	1374123	3770850	1.02e+07
disposition	2.125	.6019925	1	4
orientation	5.6	3.298586	1	10

Table 3.10: Pankrác - Summary statistics:

	Mean	St. Dev.	Min	Max
floor	4.187879	1.920962	1	8
square	66.30201	28.40613	29.26	132.78
terrace	.1212121	.3273672	0	1
enclosed_bal	.0242424	.1542691	0	1
balcony	.6363636	.4825101	0	1
price	4204141	1888315	1124230	1.06e+07
disposition	2.066667	.924816	1	4

Table 3.11: Vyhlídka - Summary statistics:

	Mean	St. Dev.	Min	Max
floor	2.313559	1.142592	1	4
square	73.12034	28.08965	33	220.3
garden	.1313559	.3385071	0	1
enclosed_bal	.4491525	.498465	0	1
price	3800334	1707008	1413461	1.21e+07
disposition	2.440678	1.08798	1	6
basement	.4915254	.5009907	0	1
y07	.3983051	.4905894	0	1
y08	.4194915	.4945247	0	1
y09	.1822034	.3868325	0	1

Appendix B: Histogram plots of residuals

Figure 4.1: Argentinská – Histogram: Plots of residuals

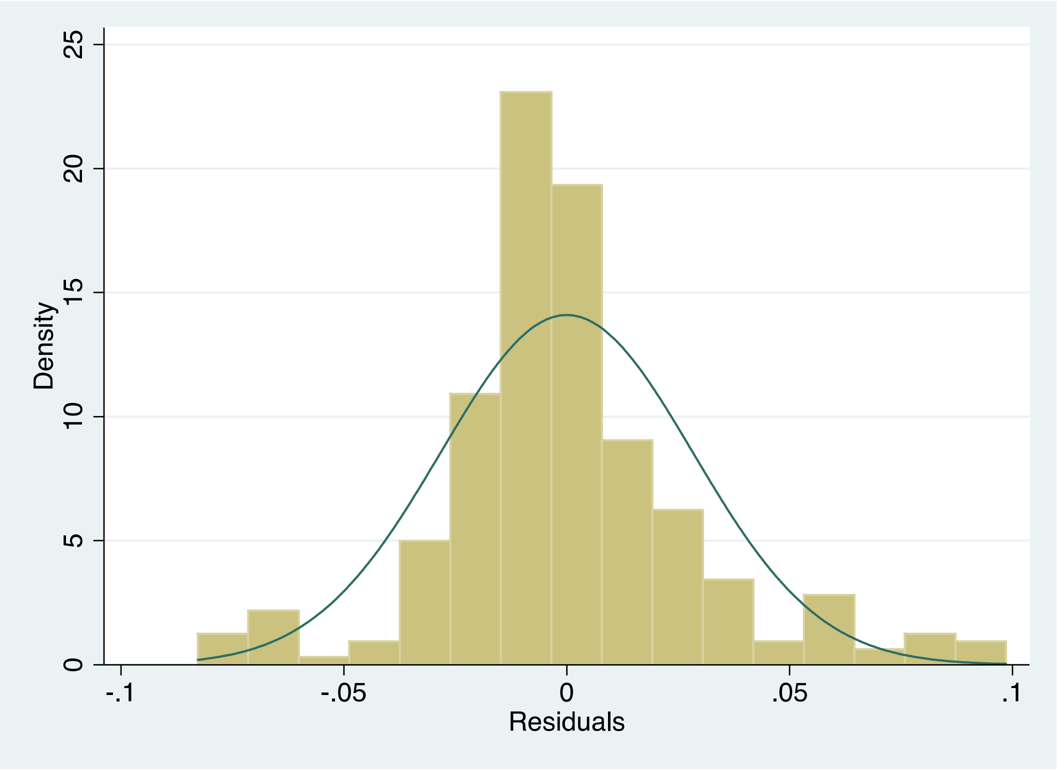


Figure 4.2: Uhříněves – Histogram: Plots of residuals

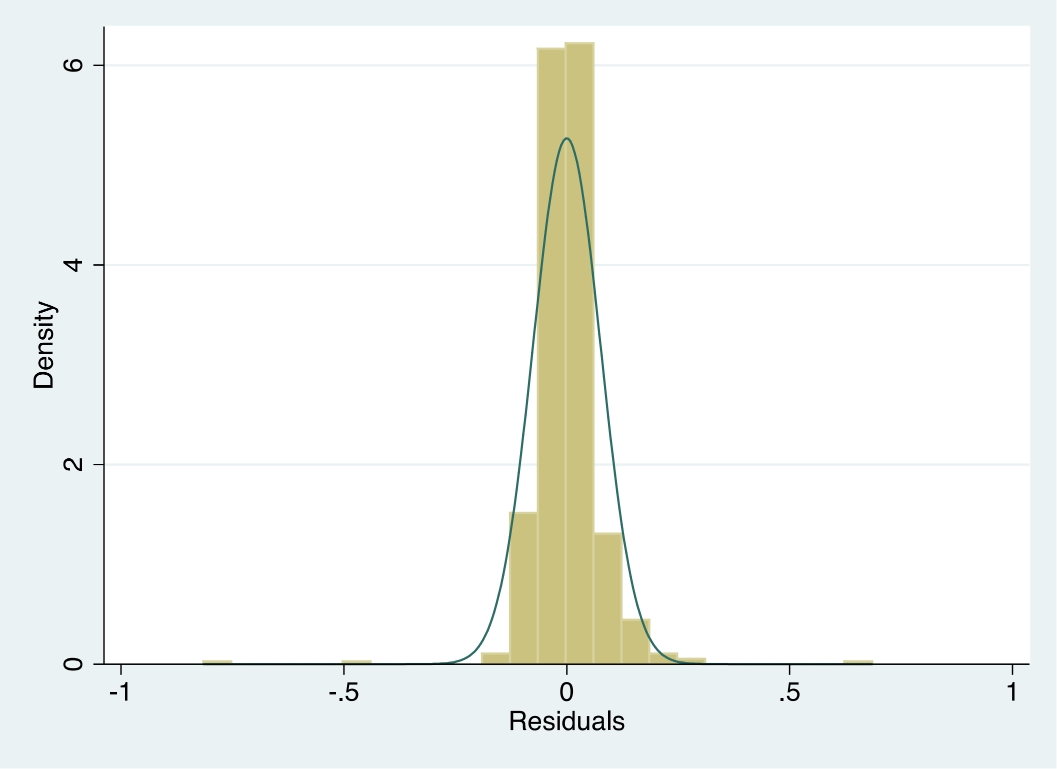


Figure 4.3: Aréna – Histogram: Plots of residuals

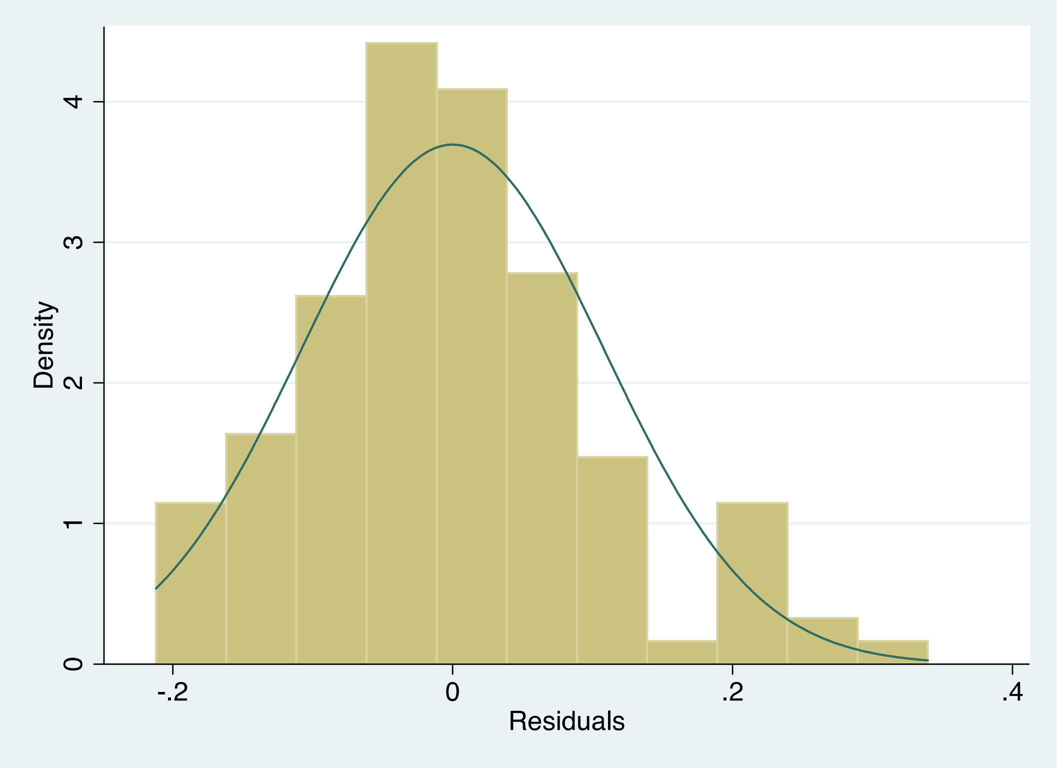


Figure 4.4: Kamýk – Histogram: Plots of residuals

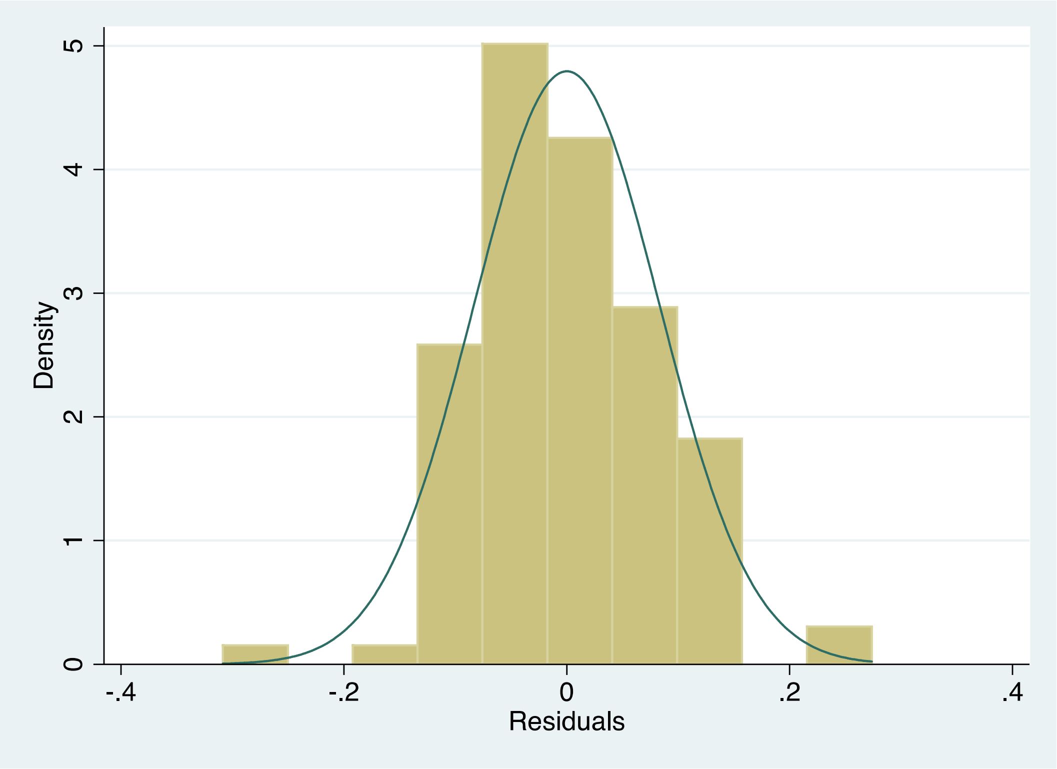


Figure 4.5: Luka – Histogram: Plots of residuals

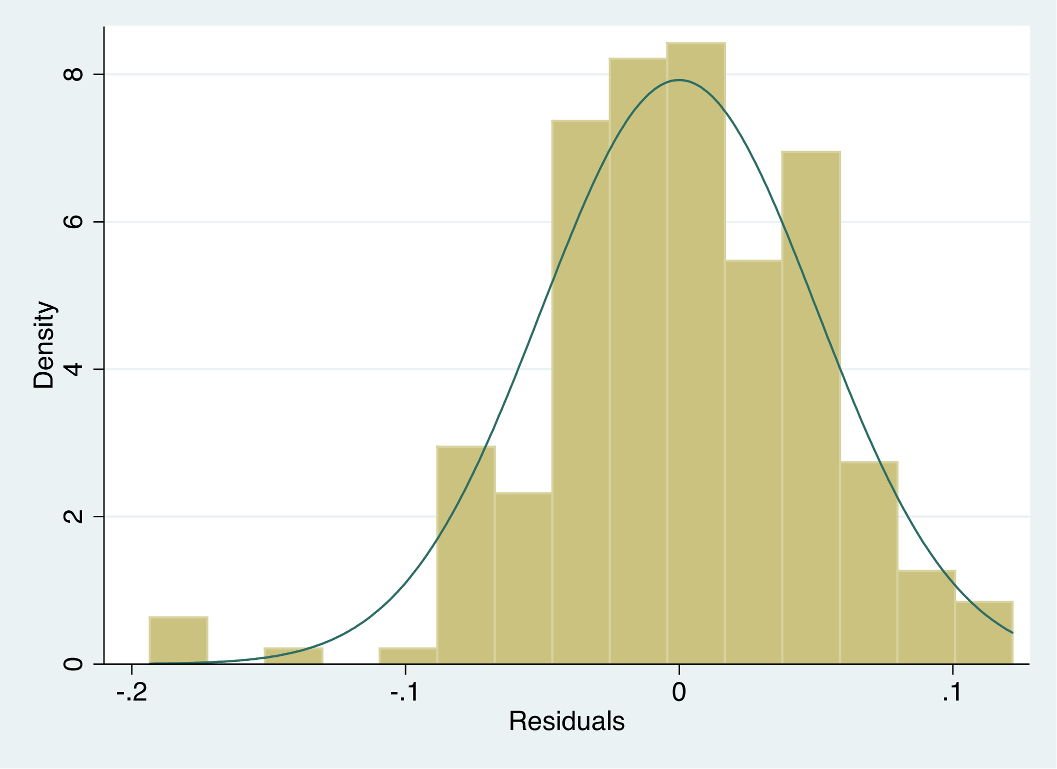


Figure 4.6: Osadní – Histogram: Plots of residuals

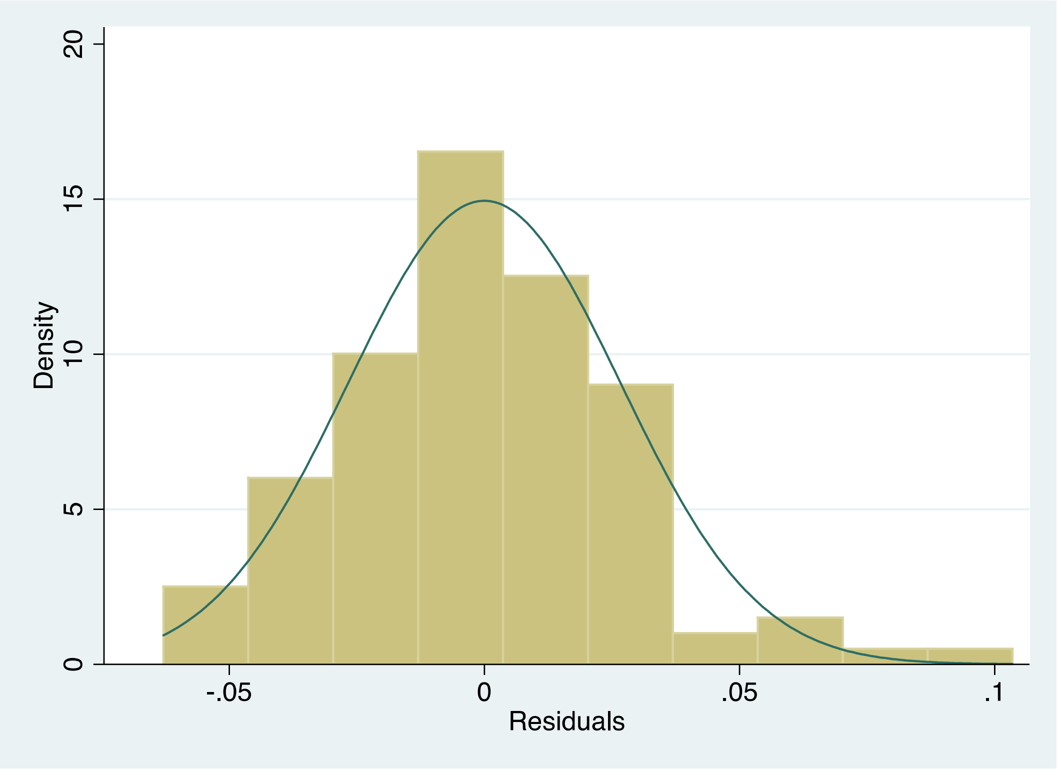


Figure 4.7: Pankrác – Histogram: Plots of residuals

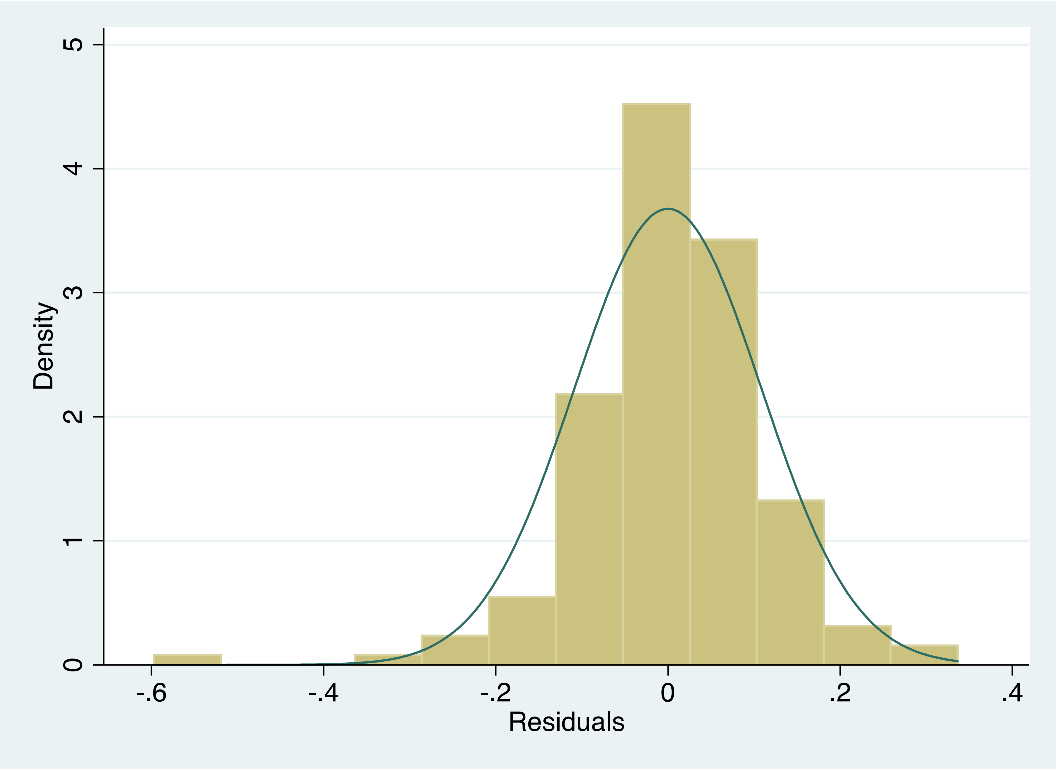
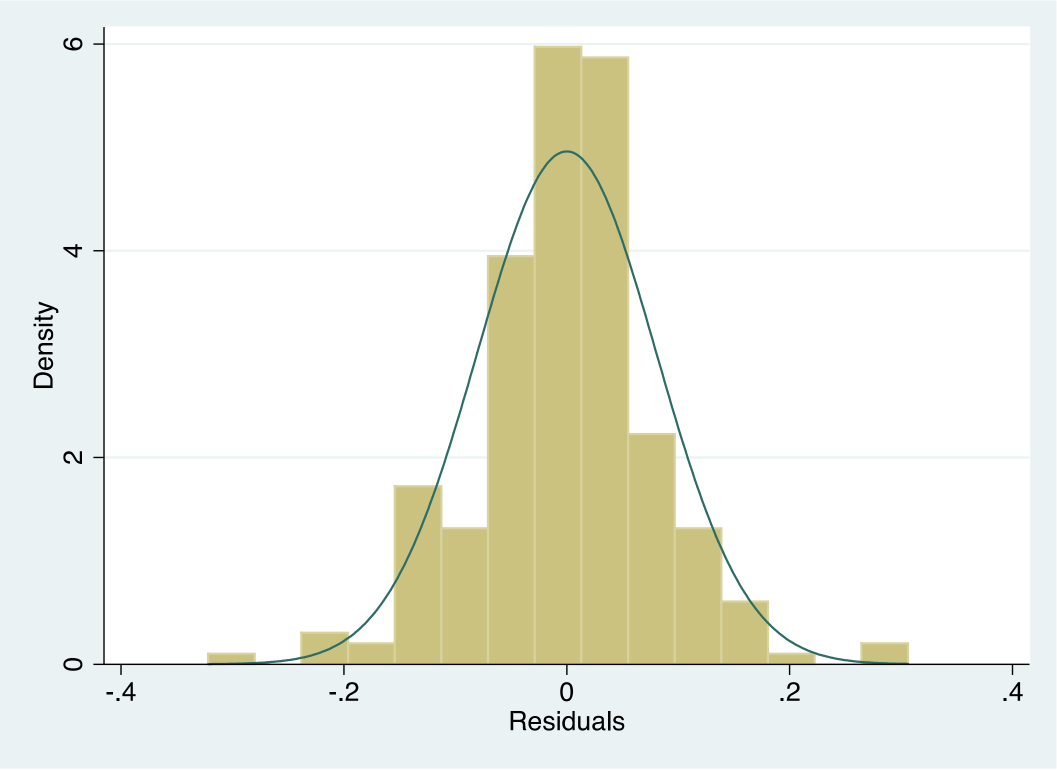


Figure 4.8: Vyhlídka – Histogram: Plots of residuals



Appendix C: Pooled cross sectional projects: calculations

The following calculations are based on versions of the following two partial derivatives:

$$\frac{\partial \text{log(price)}}{\partial \text{square}} \text{ and } \frac{\partial \text{log(price)}}{\partial \text{year}}$$

Na Vyhlídce

The averaged sized-flat is 73 square meters.

A)

$$\frac{\partial \text{log(price)}}{\partial \text{square}} = 100\% \times (0.024 - 2 \times 0.00006 \times 73 - 1 \times 0.00039 - 0 \times 0.00253)$$

$$= 1.485\%$$

After simplifying and omitting insignificant estimates, we get:

$$100\% \times (0.024 - 2 \times 0.0006 \times 73) = \mathbf{1.524}\%$$

For Year 2008 = 1, Year 2009 = 0 and square = 73

B)

$$\frac{\partial \text{log(price)}}{\partial \text{square}} = 100\% \times (0.024 - 2 \times 0.00006 \times 73 - 0 \times 0.00039 - 1 \times 0.00253)$$

$$= \mathbf{1.271}\%$$

All the estimates are significant.

For Year 2008 = 0, Year 2009 = 1 and square = 73

C)

$$\frac{\partial \text{log(price)}}{\partial \text{year}_{2008}} = 100\% \times (0.041 - 73 \times 0.00039) = 1.253$$

There are no significant estimates, so there is **no change**.

D)

$$\frac{\partial \text{log(price)}}{\partial \text{year}_{2009}} = 100\% \times (0.313 - 73 \times 0.00253) = \mathbf{12.83}\%$$

Both of the estimates are significant so the result holds.

Uhříněves

The averaged sized-flat is 56 square meters.

A)

$$\frac{\partial \log(\text{price})}{\partial \text{square}} = 100\% \times (0.033 - 2 \times 0.00015 \times 56 - 1 \times 0.00002 - 0 \times 0.00131)$$

$$= 1.618\%$$

After simplifying and omitting insignificant estimates, we get:

$$100\% \times (0.033 - 2 \times 0.00015 \times 56) = \mathbf{1.62}\%$$

For Year 2018 = 1, Year 2019 = 0 and square = 56

B)

$$\frac{\partial \log(\text{price})}{\partial \text{square}} = 100\% \times (0.033 - 2 \times 0.00015 \times 56 - 0 \times 0.00002 - 1 \times 0.00131)$$

$$= \mathbf{1.489}\%$$

All the estimates are significant.

For Year 2018 = 0, Year 2019 = 1 and square = 56

C)

$$\frac{\partial \log(\text{price})}{\partial \text{year}_{2018}} = 100\% \times (0.101 - 56 \times 0.00002) = 9.988\%$$

There is one insignificant estimate so:

$$100\% \times (0.101) = \mathbf{10.1}\%$$

D)

$$\frac{\partial \log(\text{price})}{\partial \text{year}_{2019}} = 100\% \times (0.441 - 56 \times 0.00131) = \mathbf{36.76}\%$$

Here, both of the estimates are significant so the result holds.

Appendix D: All projects: Residual analysis

Argentinská's Stata code for the residual analysis figure can be seen below. The rest of the projects' codes are similar. Baum (2006) used similar code without square^2 in his book, called An introduction to modern econometrics using Stata.

```
//PREDICTION LINE AND RESIDUALS: Argentinska
//Drawing prediction line, evaluated at means of other explanatory variables
qui reg lprice square c.square#c.square i.floor balcony i.n_dispozition, robust
qui sum square if e(sample)
scalar m_square = r(mean) // store mean into scalar
qui sum balcony if e(sample) //quietly summarize balcony
scalar m_balcony = r(mean)
qui sum 2.floor if e(sample)
scalar m2_floor = r(mean)
qui sum 3.floor if e(sample)
scalar m3_floor = r(mean)
qui sum 4.floor if e(sample)
scalar m4_floor = r(mean)
qui sum 5.floor if e(sample)
scalar m5_floor = r(mean)
qui sum 6.floor if e(sample)
scalar m6_floor = r(mean)
qui sum 7.floor if e(sample)
scalar m7_floor = r(mean)
qui sum 2.n_dispozition if e(sample)
scalar m2_dispozition = r(mean)
qui sum 3.n_dispozition if e(sample)
scalar m3_dispozition = r(mean)
qui sum 4.n_dispozition if e(sample)
scalar m4_dispozition = r(mean)

//The prediction line is below:
gen pred_means = _b[_cons] + _b[square]*square + ///
_b[c.square#c.square]*square*square + ///
_b[2.floor]*m2_floor + _b[3.floor]*m3_floor + _b[4.floor]*m4_floor + ///
_b[5.floor]*m5_floor + _b[6.floor]*m6_floor + _b[7.floor]*m7_floor + ///
_b[balcony]*m_balcony + _b[2.n_dispozition]*m2_dispozition + ///
_b[3.n_dispozition]*m3_dispozition + _b[4.n_dispozition]*m4_dispozition

//Drawing prediction line evaluated at means of all the other explanatory variables.
//Square is on the x-axis.
twoway (scatter lprice square) (line pred_means square , sort(square)), ///
title (Predicted Line Evaluated At Means) ///
legend(order(1 "Actual Data" ///
2 "Predicted lprice at means"))
graph export graph0.pdf, replace
//end
```

Figure 5.2: Uhříněves – Predicted line and actual observations

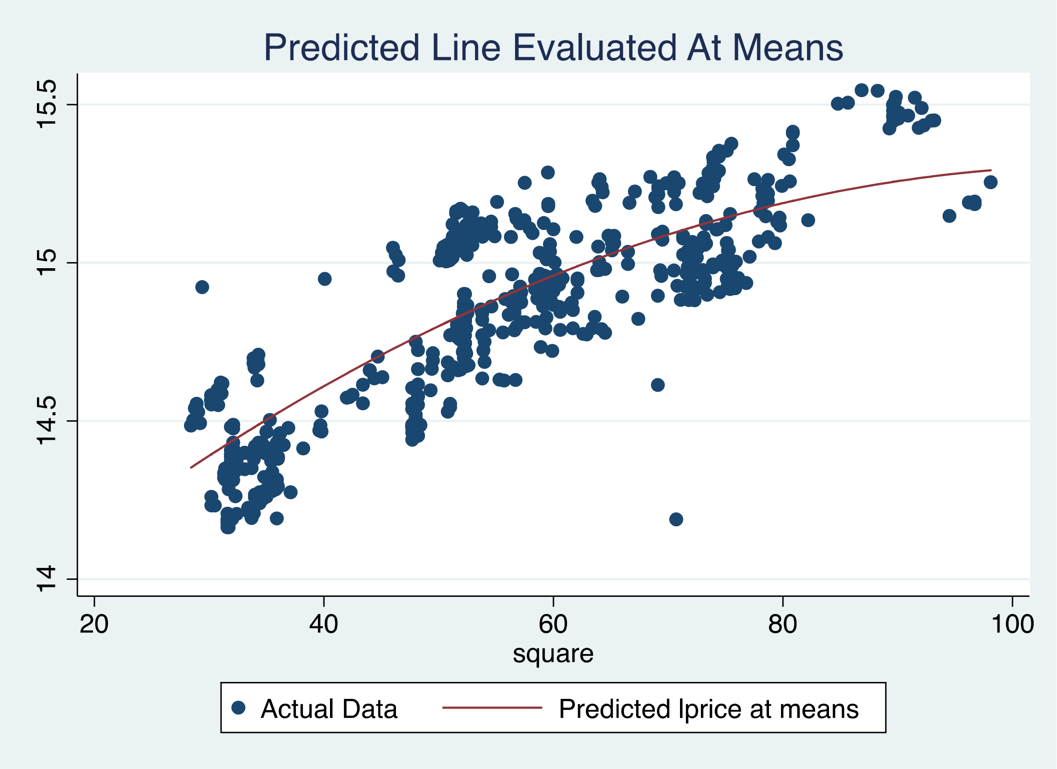


Figure 5.3: Aréna – Predicted line and actual observations

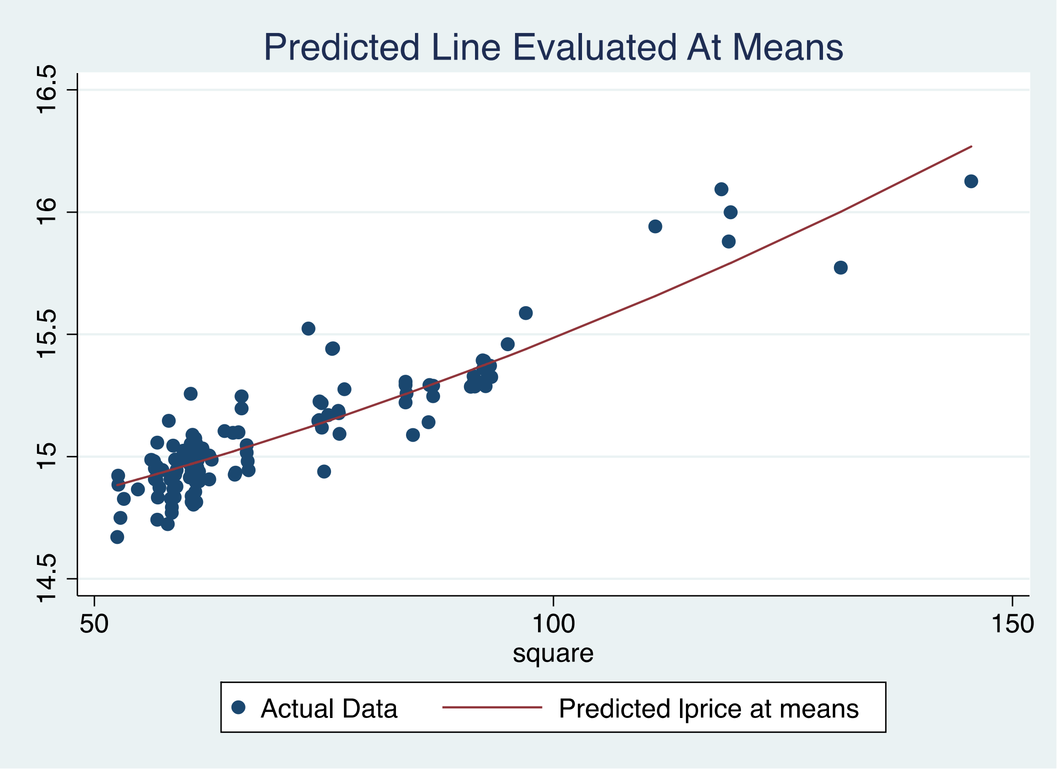


Figure 5.4: Kamýk – Predicted line and actual observations

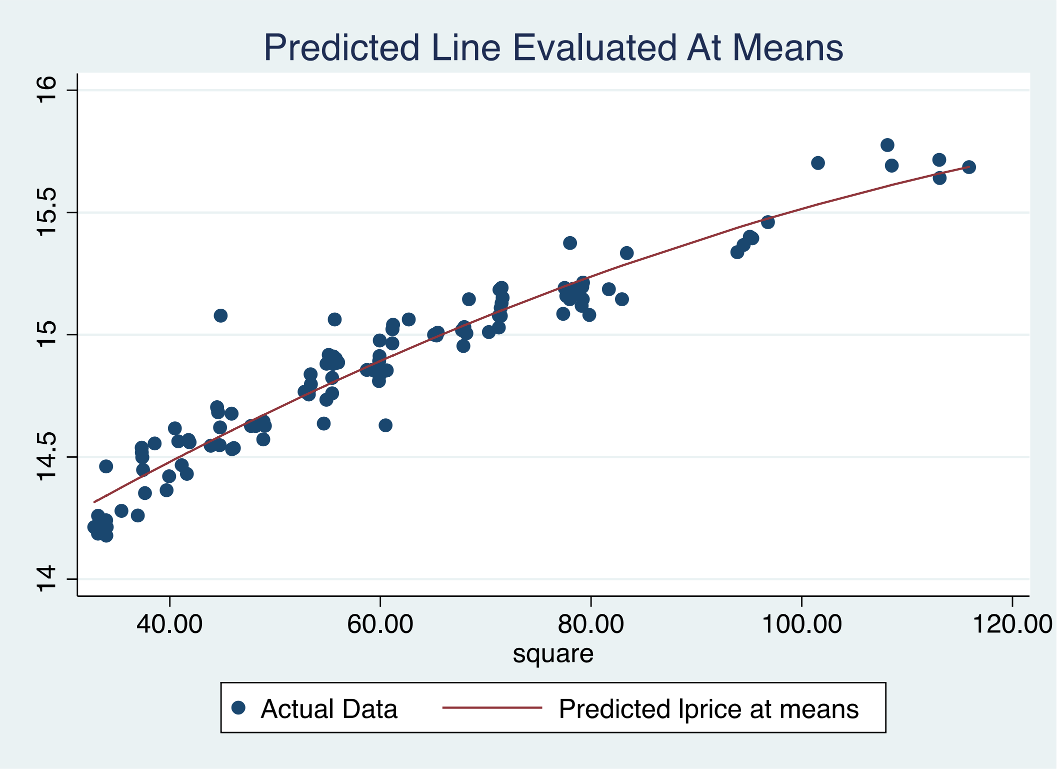


Figure 5.5: Luka – Predicted line and actual observations

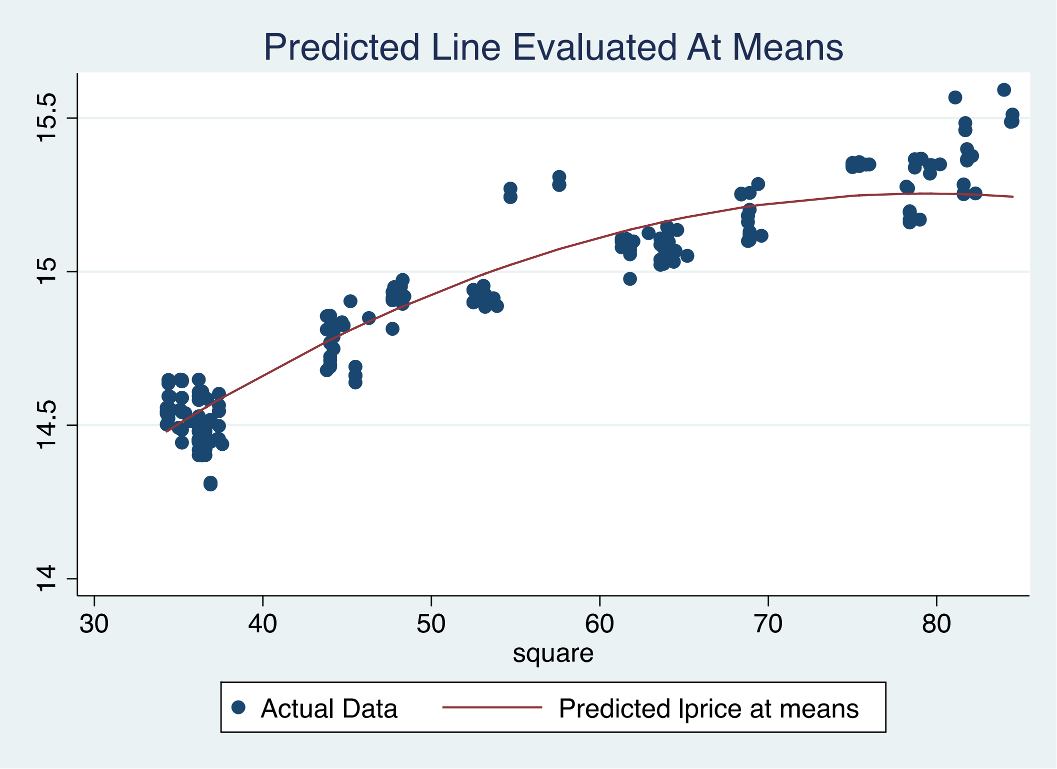


Figure 5.6: Osadní – Predicted line and actual observations

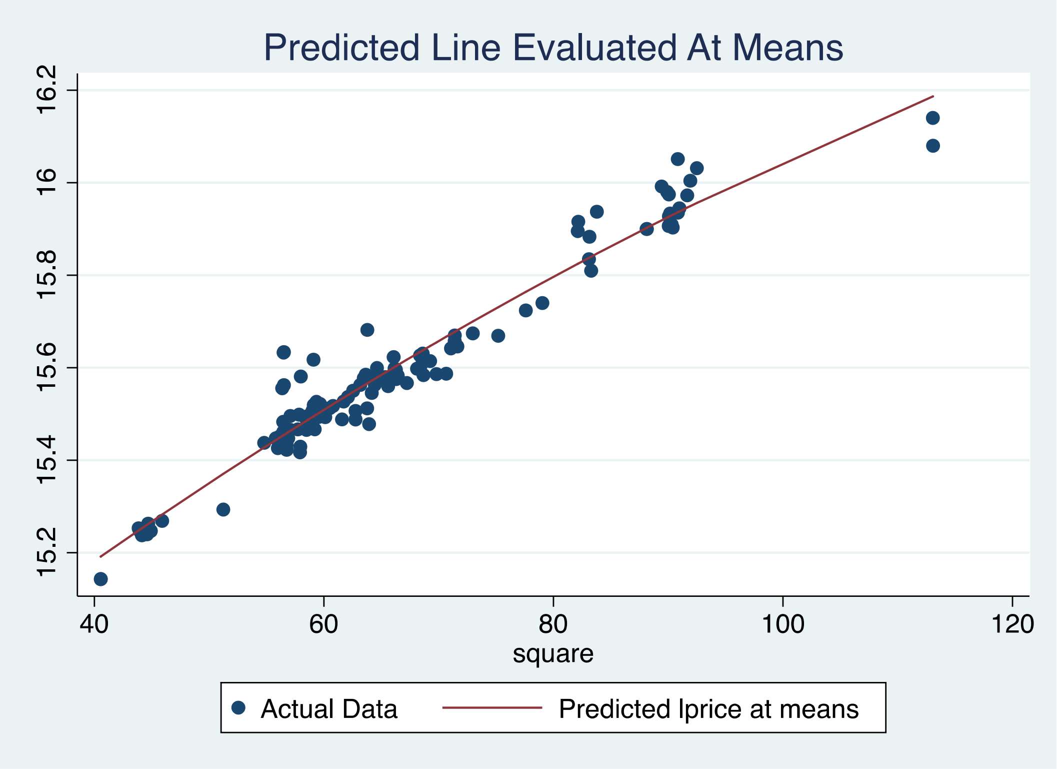


Figure 5.7: Pankrác – Predicted line and actual observations

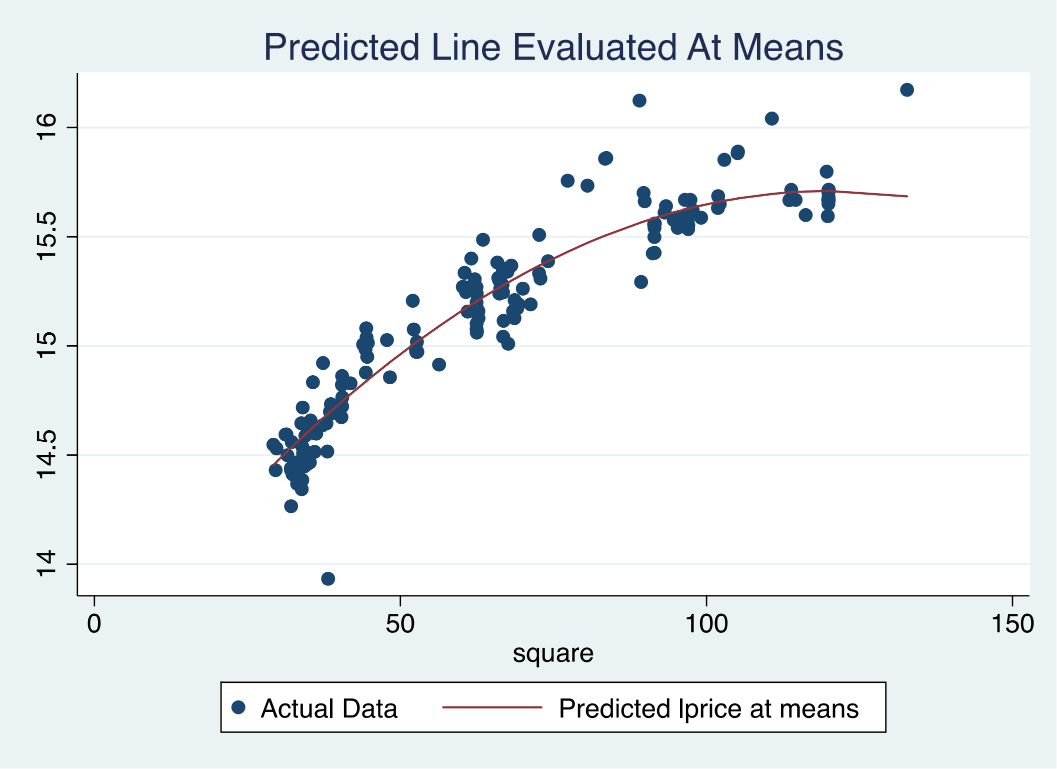
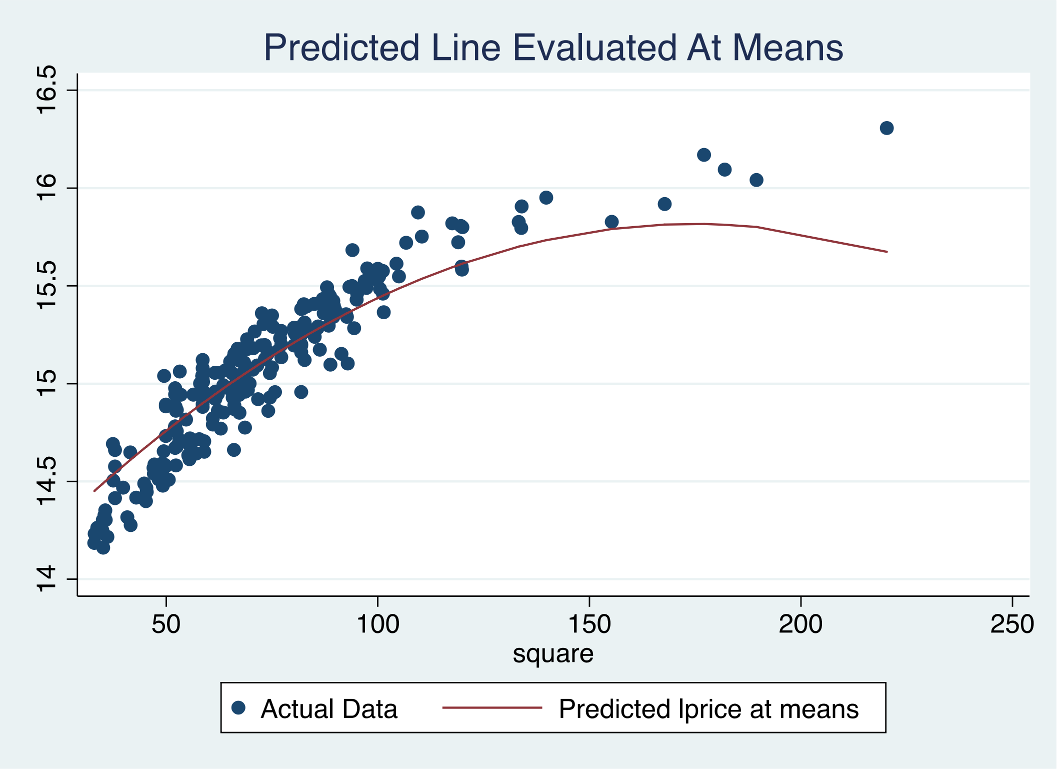


Figure 5.8: Vyhlídka – Predicted line and actual observations


Predicted Line Evaluated At Means
14
14.5
15
15.5
16
16.5
50
100
150
200
250
square
● Actual Data
Predicted lprice at means