Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science

Master Thesis

Applied Mathematics

Valuation of residential real estate in the Netherlands

Author:
A.R. Harinandansingh

October 31, 2019

Delft University of Technology

Valuation of residential real estate in the Netherlands

by

A.R. Harinandansingh

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday November 15, 2019 at 3:30 PM.

Student number:	4163559	
Project duration:	February 1, 2019 – November 15, 2019	
Thesis committee:	Dr. J. Söhl,	TU Delft, supervisor
	Prof. dr. ir. G. Jongbloed,	TU Delft
	Dr. C. Kraaikamp,	TU Delft
	S. van der Aa MSc,	TJIP/Newest Industry

This thesis is confidential and cannot be made public until January 1, 2025.

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Preface

This thesis has been written as part of the Master Applied Mathematics at the TU Delft. It is the final requirement to obtain the degree of Master of Science. The duration of the graduation project took nine months.

The thesis has been written in collaboration with TJIP B.V. and focuses on a currently hot topic: house price modeling. Due to the large price changes on the housing market, TJIP was interested in obtaining current market values of all residential properties in the Netherlands.
I would like to thank TJIP for the opportunity to write my thesis about this interesting topic.

Furthermore, I want to thank the people I have worked with the most during this project. From TU Delft, I want to thank my supervisor Jakob Söhl for the guidance. His help and feedback were really useful. And from TJIP I want to thank my supervisor there, Sander van der Aa, but also Floor van de Merbel for their input and suggestions.

Finally, I would like to thank Geurt Jongbloed and Cornelis Kraaikamp for being part of the graduation committee.

A.R. Harinandansingh
Delft, October 2019

Contents

1 Introduction

What is the actual value of a house and which factors contribute the most to it? In this thesis we do a thorough research and try to come up with an answer. We will set up models that approximate the current market values for all houses in the Netherlands. To do this, we use a lot of data from different sources. Various characteristics of houses, the location and the trend on the housing market will be used.

This thesis has been written in collaboration with TJIP B.V. located in Delft. They call them self as 'platform engineers' and 'platform investors' and develop business platforms where people and businesses can interact smarter and work more efficient. This graduation project is for one of the platforms of TJIP that focuses on the housing market. On this platform potential buyers and sellers can interact. Every property in the Netherlands is available with a lot of addition information. For example, there is data about the living space, construction year, nearest distance to schools and supermarkets, maintenance costs and property taxes. By signing up, users can make an account and claim their house. Registered users can approach each other and make the first contact with the purpose to reach an agreement for a possible house transaction. By working more efficiently, they can choose to avoid the intervention of a real estate agent and go straight away to a notary after possible mortgage requests are approved.
Besides the data examples we mentioned, there is also an estimate given of the current value of a house. And that value is where this thesis is all about. This value can be used as starting point in the negotiations to reach an agreement.

The goal is to come up with an estimate that approximates the current value of a house. The question is what this value is and where it is based on. Is it based on only the characteristics of a property like the size of the living space and the type (e.g. apartments, detached houses) or are other factors involved as well? Is it determined by supply and demand on the housing market? Does this value even exists? Actually, a value cannot be determined accurately in terms of euros. It will always be an estimate and therefore be debatable.

1.1. Literature study

To gain knowledge into this subject, we start with a literature study. We look for several scientific papers with a strong mathematical component that focuses on this subject. We are interested in which approaches already exists and where we can make a difference. We list the papers that we have used for this literature study below and give a short description about their approach.

There are two methods that are often used for real estate pricing: repeat sales models and hedonic pricing models. We use [13] to obtain some basic information about the repeat sales method and the developments over the years and [14] for information about hedonic models. Note that [13] and [14] are different chapters from the same book.

- In 1963 the repeat sales method was proposed by Bailey, Muth and Nourse [15]. Basically, in a repeat sales model only houses which are sold more than once are used. The sale prices, the date of sale and the corresponding address of a property are used to set up a regression model. It enables us to come up with a price index and obtain information about the price development. After 1963, variations on this method were proposed to construct price indices. For instance, the Standard and Poor's/Case-Shiller home price indices in the US or repeat sales indices for cities in the United Kingdom by the UK Land Registry.

The original Case and Shiller model (1987 [16], 1989 [17]) expands the one of Bailey, Muth and Nourse by accounting for heteroscedasticity due to the gap time between sales. Case and Shiller proposed a weighted least squares approach to correct for a non-constant variance. The weights are derived by regressing the squared residuals from the standard ordinary least-squares repeat sales regression on an intercept and the time interval between sales.

Even though we only need data regarding the sale price, time of sale and the location, a repeat sales model has several disadvantages. For instance, useful transaction data will be lost because only data of properties that are sold more than once is considered. Also, it does not distinguish between the several property types. Furthermore, a sample where only properties that are sold more than once are considered does not represent the real estate market well. In general, house sales occur infrequently and only a small percentage is actually sold in a certain time scale. They will probably not be sold for the same price. For instance, an investor can buy a property with the purpose of making a financial profit, possibly by doing some renovations. These renovations are captured into the increased asking price. However, using this increased price for estimates of other properties that are not sold more than once will lead to bias, since it is not based on economic and/or real estate market effects.

- In a hedonic model we assume that the value of property is determined by different characteristics with different degrees in terms of their contribution. Regression techniques can be used to estimate these contributions, which are known as the regression coefficients in a regression setting. The characteristics can be for instance related to the property (e.g. type, size of the living space) or to the location (e.g. neighborhood, living conditions). Usually, prices are computed for different time periods. This is done by solving a regression equation for each time period seperately. The reason is that over time the regression coefficients can change and thus are not necessarily constant. From these predictions for each time period, a price index can be constructed.

- Next we look at a paper that extends the repeat sales method: [18]. In this paper the objective is to develop a practical model to predict prices from which a price index can be constructed. It considers information regarding the sale price, time of sale and location. The proposed model applies the repeat sales idea in a new way by taking into account some critical points. For instance, not only repeat sales are considered, but also single sales. The latter includes new home sales which are usually more expensive and bring new pricing trends to the market.

  Sale prices become less useful the longer it has been since the last sale. Therefore, an underlying first-order autoregressive time series is used in [18]. Regarding the data in [18], the authors analyze single-family home sales for twenty US metropolitan areas from July 1985 through September 2004. They show that their model has better predictive abilities than the benchmark S&P/Case–Shiller model.

  An outline of the autoregressive model proposed in [18], together with the repeat sales models by Bailey, Muth and Nurse, Case and Shiller and the Standard and Poor's (S&P)/Case-Shiller model can be found in [19]. These models are all compared in [19] by looking at the predictive power.

- The next paper is written by two employees of the Swiss National Bank and received scientific support by the National Center of Competence in Research "Financial Valuation and Risk Management" [20]. In this paper the proposed models are not only based on the observed dynamic on the real estate market. It states that the fundamental price is the sum of the discounted future period costs that arise from owning a house. The expected capital gain in period t is derived via the expected house price in period $t+1$.
  For their model the writers consider the user costs; the costs that arise from owning a house for one period. They consider several factors: the mortgage the owner of a house has to pay, a fixed fraction of the value of the house (i.e. maintenance costs, constant property taxes, and a constant risk premium) and the expected capital gain. They also come up with a rent model, since this affect the factors. In the end this leads to fundamental price equation. By using a vector autoregression (VAR) model, future values

of the price can be estimated. To judge the model the writers use data from countries with different developments of the housing market: the United States, the United Kingdom, Japan, Switzerland and the Netherlands.

- A paper by researches of TU Delft about house prices in the Netherlands is [21]. In particular, the researchers take into account the regulated market of the Netherlands for their analysis. They investigate the long-run house-price relations in the Netherlands from 1982 to 2008 with the help of some statistical theory. The authors mention that The Netherlands has a predominant social housing rental sector, a strongly subsidized housing market, and a highly inelastic supply sector. These together contribute to the special price path in the Dutch market.

  For the average house price, their model uses factors like inflation rate, the household income and the mortgage interest rate. In the long-run, incomes and interest rates function as the two prime forces driving price dynamics, whereas the role of inflation is limited.
  Quarterly data from the second quarter of 1982 to the first quarter of 2008 is used in this paper. Therefore, the researchers get insight into the price development for a 26-year time period.

- Finally, we have a look at [22]. This is a more economics oriented paper that focuses on the sustainability of house prices. This is in our case rather an informative paper regarding house prices than one with a useful modelling approach. It considers factors like household income, demand and stock that affect long-run trends in house prices.

Even though not all papers were based on the Dutch real estate market, it has given us an idea about real estate price modelling. As we can see, the researchers in the papers mentioned above have constructed models based on a lot of different data over a large time span which is probably not all open data. Their research may or may not be funded. Most of the data they have, we have not. And this is essential in deciding which method(s) we can use. The information from the papers we take with us, but our models will be different.

1.2. Problem approach

Ideally, sale and transaction prices would help us a lot in this project. It can be used for valuation of nearby properties. Transaction prices of houses in the Netherlands are registered by the Dutch cadastre [23] (Land Registry), which is a governmental institution. Unfortunately, this data is not available for us, unless we pay for it. Because of that, we need to be a bit creative and come up with alternative approaches.

To start we need some reference values of houses. For this we use WOZ values. In Dutch, WOZ is an abbreviation for 'waardering onroerende zaken', which can be translated as the valuation of real estate. The WOZ law regulates the determination of the WOZ values. These values are estimated by municipalities on January 1 and are revealed one year later. In particular, we use WOZ values of 2017. So these WOZ values are measured on January 1 of 2016. Property owners receive the WOZ value of their property usually in January or February. The WOZ value is used for tax purposes. To determine the WOZ value of a property, municipalities compare it with sales prices of recently sold properties in the neighborhood. The WOZ value is then an estimate of the sale price if the property was sold on the reference date (January 1 of the previous year). More details about the WOZ value can be found via [9] and [10].

We have access to most WOZ values of 2017, but not all of them. We first set up a model that can predict this WOZ value. Therefore, we can obtain an estimate of the WOZ for the missing properties. We do this first for only one municipality which will be Delft. When this is done, we scale up the region and use the model on provincial level where we will start with Zuid-Holland. Subsequently, other provinces are next where sometimes provinces will be put together in one model. After we have an estimate of the WOZ for all properties, so including the properties for which we initially have a WOZ value, we try to find a relationship with asking prices of houses that has been put on sale. This enables us to predict the asking price. From these predictions we can derive estimates of the actual values.

We do not want to use the original WOZ values when searching for a relationship with the asking prices. Instead, we want to give our own input, which is derived from a model, to the final predictions. Therefore, there is a certain structure and coherence in our predictions. Furthermore, we are also capable of detecting possible outlying and incorrect WOZ values estimated by the municipalities.

1.3. Thesis structure

In Chapter 2 we discuss everything related to the data. We give an overview of what we have and describe the whole data cleaning process. In Chapter 3 we discuss the mathematical theory and methods that we will use. After that we apply the theory to the municipality Delft and set up our first model. This is described in Chapter 4. Hereafter, we scale up to provincial level. Details can be found in Chapter 5. Next we try to find the relationship between predicted WOZ values and asking prices. This is discussed in Chapter 6. We finish with a discussion, our conclusion and some recommendations in Chapter 7. In the Appendix some background information is given about the models. Details about this will be mentioned in the chapters.

2
The data

In this chapter we discuss the data we have. We list all variables and give some additional information if we consider it necessary. In the first section we start with an overview of the data for Delft since this is the municipality we use for our first model. After that we give information about the data on provincial level and data about the asking prices. The second section is devoted to the whole data cleaning process. In the last section we explain why some data that initially could be interesting for this project has not been used.

2.1. Data for our models

2.1.1. Data for Delft

Various data we have for every combination of postal code and house number for the municipality Delft. We know for example the size of the living space and in which district and neighborhood the dwelling is located. We give an overview of the data we have for a dwelling.

- Size of the living space in m^2
- Construction year
- The district where the dwelling is located.
- The neighborhood where the dwelling is located.
- Energylabel
- WOZ-value
- The closest distance (in meter) to 8 different facilities: school, supermarket, hospital, highway ramp, bus station, shopping mall, train station and residential boulevard.
- The type of a property. Here we use the same 5 categories as the Dutch cadastre (Land Registry) [2].
  1. Apartment
  2. Terraced house
  3. House located on a corner or at the end of a terraced house construction.
  4. Semi-detached house
  5. Detached house

  In our models only one type in the list above can be assigned to a property. It is possible for example that a semi-detached house is located on a corner. In that case it will we ranked under semi-detached houses. The list is, in general, in ascending order in terms of value.
- Leefbaarometer. It can be defined as the extent to which the living environment meets the conditions and needs that are imposed on it by humans. This measures the quality of life in district and neighborhoods and is based on the five so-called dimensions: habitants, dwellings, facilities, safety and physical environment. It is expressed as a numerical deviation (positive or negative) from the national average and is published every two years. We use the data of 2016.

Note that for a property or neighborhood sometimes data can be missing. For example, there is no energylabel available or the leefbaarometer data is missing. This is a general problem in our models, which we discuss later.

2.1.2. Data for the provinces

In the previous chapter we mentioned that we start with a model for Delft and after that we will model on provincial level. By scaling up the location, we encounter new problems. Data sets become larger and the computation time increases. In the Netherlands, a municipality can have more than one city. With city we mean here for example what we write on the envelope when sending a letter via non-digital post. In the first model, we work with the municipality Delft. This municipality has only one city, with the same name. On provincial level, we have municipalities with more than one city. For example, the municipality Nieuwkoop in the province Zuid-Holland has seven different cities: Nieuwkoop, Nieuwveen, Noorden, Ter Aar, Vrouwenakker, Woerdense Verlaat and Zevenhoven.

The data that we have for the municipality Delft, we have for the properties in all 12 provinces in the Netherlands as well. Furthermore, we know in which municipality and in which city the property is located.

2.1.3. Asking prices

We have access to asking prices of houses that has been put on sale in the last 2.5 years. We have information about the address, the date when a house was put on the housing market and the asking price. Sometimes a house appears more than once in the data set. This means it was put on the market again on another date. Additionally, the price can be different as well. If this happens in a small period of time, it probably means the asking price was reduced due to low interest.

2.2. Data cleaning

When we obtain our data it is in raw form and cannot be used for statistical analysis. The whole process of transforming the raw data into a data set that can be used for further analysis is called data cleaning. In general, data cleaning is more time consuming than the statistical analysis itself.
Our data comes from different resources. Examples are the BAG (Basisregistratie Adressen en Gebouwen) [25], RVO (Rijksdienst voor Ondernemend Nederland) [26] and the Leefbaarometer [27], which is part of the Dutch Ministry of the Interior and Kingdom Relations. The biggest challenge therefore is to merge all data sets to obtain one new set that provides all information. In this section we describe our data cleaning process.

First data set

The first data set we obtain via TJIP and is from [32]. This set provides us with information about postal codes, house numbers and possible house number additions. These house number additions can have letters, numbers or both. Here is an example from our data set.

number_complete	number	number_letter	number_addition
11	11	NA	NA
11A	11	A	NA
11B	11	B	NA
12	12	NA	NA
12A	12	A	NA
12A-1	12	A	1
12B	12	B	NA
13	13	NA	NA
13A	13	A	NA

Figure 2.1: Structure of house numbers and number additions of properties in the Netherlands.

In the first column we have the complete house number, which is a composition of the next three columns.

Letters in the column ‘number_letter’ are put right behind the house number. Elements in the column ‘number_addition’ are put behind the element in the column ‘number_letter’ (or the column ‘number’ if there is no letter addition) after a minus sign. Although not displayed here, the last column can have letters as well. Furthermore, we have the city, district, neighborhood, size of the living space, construction year, nearest distances to eight different facilities, and some energylabels and property types. We call this data set as our original set and want to supplement it with data from other sources.

BAG data

The next step is to use the data set from the BAG to obtain the purpose of use of the property. We are only interested in properties for residential use. However, there can also be sport halls, offices, shopping centres and properties for educational purposes. We want to discard everything that is not for residential use. We distinguish properties by an object_id. This is the complete house number pasted behind the postal code. This allows us to match the BAG data with our data set. After that we know for all properties whether it is for residential use.

Energylabels

As mentioned before, we only have access to a small amount of energylabels. More energylabels can be obtained via RVO. They have different data sets for definitive and temporary energylabels. Energylabels show how sustainable the energy performance of a property is. Property owners are obliged to have a definitive energylabel when they sell or rent their property. This can be requested via governmental agencies. The data set with temporary labels is an estimation of the labels based on different factors. More about this can be found via RVO.

The data set with definitive labels contains energylabels for 3.8 million properties. The data set with temporary labels has energylabels for almost 8 million properties. These to numbers together exceed the total number of properties in the Netherlands which is around 9 million. It is possible that a property is contained in both data sets. As mentioned above, a definitive label becomes available when a property is sold or when a label is requested by the property owner. Subsequently, it can be included in the data set with definitive energylabels. However, the property is not removed from the data set with temporary energylabels. We always start with matching the definitive energylabels. After that we fill up the missing labels by using the data set with temporary labels.

As we did with the BAG data, we now want to match the data from RVO on the object_id as well. Unfortunately, the data sets of RVO do not have the house number separation layout as in Figure 2.1. In the data sets of RVO all number additions are taken together. Therefore we do not know whether we have to take a minus sign as separator. Another option is to match on a source_id. This is an integer of 14 or 15 numbers that classifies a property. This integer is only available in the RVO data set for definitive labels and the BAG data set. By using the BAG we are able to get the object_id of the properties in the RVO and can still match it with the properties in our original data set. However, the source_id is not available for all properties in the RVO data set. To still incorporate the labels, we take the number addition and read out the first sign. If this is a letter, it is most often a letter addition (as in the third column in Figure 2.1) and we put it right behind the house number. If not, we separate it by a minus sign behind the house number. By doing this, we obtain the complete house number (as in the first column in Figure 2.1) and can compose the object_id. Therefore we can still match these remaining labels.

For the temporary labels we do not have a source_id and compose the object_id by reading out the first element of the number addition. In general, only a small amount of properties have a letter addition or even a number addition as in the last column of Figure 2.1. And if there is an addition, properties in a street with same numbers, but different additions are most often of the same type and have therefore likely the same energylabel. Because of that, we believe that little can go wrong. Besides, with the huge amount of data we have to deal with, it is almost impossible to have only correct data.

Property types

As with energylabels, we have also only a small number of property types. In the first paragraph we said that we use the same categories as the Dutch cadastre. Unfortunately, not all our data resources use these categories.

Our original set distinguish between a main type and subtype of the object. Examples of main types are canal houses, townhouses, penthouses, upstairs and downstairs apartments. Examples of subtypes are terraced houses, semi-detached houses and corridor flats. The data set with definitive labels from RVO contains property types as well and also use another typing. It uses main types like porchs, corridors and flats and a lot of eponymous subtypes. So they distinguish between main types and subtypes as well, but use another classification. This means we have to retype all properties from these two sets to assign it to one of the categories of the Dutch cadastre. It is also possible that we have for a property types from both sets that may or may not lead to the same cadastre category. Our method is to filter out first the semi-detached houses and detached houses. After that we take out terraced houses and look most often at the subtype if it is located on a corner or if it is at the end of a terraced house construction. Next we can classify most apartments. For what remains, we decide in which category a property can be placed best.

The data set with temporary energylabels contains property types as well. As with energylabels, these types are also an estimate. The vast majority of types come from this set. Fortunately, this set uses the Dutch cadastre categories with an additional category for maisonettes. This category we scale under apartments.

Leefbaarometer

Leefbaarometer on neighborhood level is given by neighborhood numbers. These numbers are derived from Statistics Netherlands. In our original data set, we only have the neighborhood names. This means that we first need to merge the leefbaarometer data with the set of Statistics Netherlands (which contains neighborhood names) by matching on neighborhood numbers. After this operation we can incorporate the leefbaarometer data into our original set by matching on neighborhood names.

WOZ values

What remains are the WOZ values of 2017. These are provided at TJIP as well and are derived from [9]. This set has the house number layout as in Figure 2.1 and is therefore easy to incorporate by matching on the object_id.

Asking prices

This data is also provided by TJIP and is from [32]. We will use this data in the end to come up with a prediction for the asking price. Regarding the data cleaning we encounter difficulties to obtain the object_id. Street and house number are given in one string. The postal code and the city are given together in another string. The postal codes we can get by reading out the first seven symbols (four numbers, a space and two letters), since the postal codes always come up first in the string.

Regarding the house numbers, it is trickier to obtain them, especially when there is a house number addition. The reason is that house numbers additions are usually separated from the house number. It takes a lot of effort to take into account all possible variations; we can have house number additions that may or may not include numerical additions, street names with more than one word, numerical symbols in street names. These are just some examples. We are capable of obtaining almost all house numbers. However, it goes not always well for number additions. If that is the case, we drop the number additions.
Furthermore, we have also newly build properties in this data set. These are usually characterized by a construction number. These properties thus have no official address yet and are dropped from the further analysis.

Furthermore, sometimes the date when a house has been put on the market is missing. If so, then it will be excluded from the further analysis as well.

2.3. Other data

As we did with the literature research to explore which methods exists, we also did a data research to find out what was available and useful for this project. All above mentioned data survived the selection. However, other data sets were available as well, but were not used for various reasons. In particular, Statistics Netherlands has a lot of open data [33]. For example, demographic data (e.g. life expectancy, birth rate, death rate, composition of the population) could be involved in this project. Or data about the average sale price of houses in a municipality per year or the price development. The reason for not using it in our models was usually that it was not available on smaller scale (e.g. cities, districts, neighborhoods) or that it could not be matched with our original data set. As we explained in this section, we want to assign all our data to a specific address (the object_id) which was not possible. Also we found data sets that we did not qualify as useful for this project. Besides, with all the data mentioned earlier in this chapter, we already have a large number of variables.

3
Regression analysis

In this chapter we discuss everything about regression analysis that is important for us. As general reference for regression we have used [1]. However, the more sophisticated computations and derivations in the first and last section we have done by ourself.

3.1. Linear regression

In a linear regression model we seek for a connection between a response variable and one or more explanatory variables. In case of one explanatory variable we have a simple linear regression model and this looks like

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{3.1}$$

where Y_i and X_i are the values of the i-th response variable and explanatory, respectively. α and β are the regression parameters, where α in particular is called the intercept. ϵ_i is an error term and is called the residual. When the relationship between the response variable and the explanatory variable appears to be linear, we can fit a simple linear regression model to our data. However, no straight line can pass perfectly through our data points. Let $\hat{Y}_i = A + BX_i$ be the fitted value by the regression model for observation i. The difference between the value and the fitted value of observation i is called the residual.

$$Y_i - \hat{Y}_i = \epsilon_i \tag{3.2}$$

Ideally, we want the residuals to be as close to zero as possible. This we can obtain by using a linear least-squares approach. This requires us to find the regression parameters of the fitted values (A and B) such that the sum of squared residuals $\sum \epsilon_i^2$ is minimized. We call A and B the least-squares coefficients. We can write

$$S(A,B) = \sum_{i=1}^{n} (\epsilon_i)^2 = \sum_{i=1}^{n} (Y_i - A - BX_i)^2 \tag{3.3}$$

where n is the number of observations. Minimizing $S(A,B)$ requires us to find the partial derivatives with respect to A and B.

$$\frac{\partial S(A,B)}{\partial A} = -2 \sum_{i=1}^{n} (Y_i - A - BX_i) \tag{3.4}$$

$$\frac{\partial S(A,B)}{\partial B} = -2 \sum_{i=1}^{n} (Y_i - A - BX_i) \cdot (X_i) \tag{3.5}$$

Setting the partial derivatives equal to zero and solve gives for A:

$$
\begin{aligned}
-2\sum_{i=1}^{n}(Y_i - A - BX_i) = 0 \\
\Longrightarrow n\cdot A = \sum_{i=1}^{n}(Y_i - BX_i) \\
\Longrightarrow A = \frac{1}{n}\sum_{i=1}^{n}(Y_i - BX_i) = \bar{Y} - B\bar{X}
\end{aligned}
\tag{3.6}
$$

And for B:

$$
\begin{aligned}
&-2\sum_{i=1}^{n}(Y_i - A - BX_i)\cdot(X_i) = 0 \\
&\Longrightarrow \sum_{i=1}^{n}(Y_i - A - BX_i)\cdot(X_i) = 0 \\
&\Longrightarrow \sum_{i=1}^{n} Y_i\cdot(X_i) - A\cdot(X_i) - BX_i^2 = 0 \\
&\Longrightarrow \sum_{i=1}^{n} Y_i\cdot(X_i) - (\bar{Y} - B\bar{X})\cdot(X_i) - BX_i^2 = 0 \\
&\Longrightarrow \sum_{i=1}^{n} Y_iX_i - \bar{Y}X_i + B\bar{X}X_i - BX_i^2 = 0 \\
&\Longrightarrow \sum_{i=1}^{n}(Y_i - \bar{Y})X_i + B(\bar{X} - X_i)X_i = 0 \\
&\Longrightarrow \sum_{i=1}^{n}(Y_i - \bar{Y})X_i = -B\sum_{i=1}^{n}(\bar{X} - X_i)X_i \\
&\Longrightarrow B = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})X_i}{\sum_{i=1}^{n}(X_i - \bar{X})X_i} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\cdot(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\end{aligned}
\tag{3.7}
$$

What rests now is to show that A and B minimize $S(A,B)$. This can be done by looking at the second order partial derivatives of $S(A,B)$ with respect to A and/or B and by looking at the Hessian matrix. This matrix is defined as

$$
H = \begin{bmatrix} \frac{\partial^2 S(A,B)}{\partial A^2} & \frac{\partial^2 S(A,B)}{\partial A\,\partial B} \\ \frac{\partial^2 S(A,B)}{\partial B\,\partial A} & \frac{\partial^2 S(A,B)}{\partial B^2} \end{bmatrix}. \tag{3.8}
$$

First we compute the second order partial derivatives.

$$
\frac{\partial^2 S(A,B)}{\partial A^2} = \sum_{i=1}^{n} 2 = 2n \tag{3.9}
$$

$$
\frac{\partial^2 S(A,B)}{\partial B^2} = \sum_{i=1}^{n} 2X_i^2 \tag{3.10}
$$

$$
\frac{\partial^2 S(A,B)}{\partial A\,\partial B} = \sum_{i=1}^{n} 2X_i \tag{3.11}
$$

Note that

$$
\frac{\partial^2 S(A,B)}{\partial A\,\partial B} = \frac{\partial^2 S(A,B)}{\partial B\,\partial A}.
$$

If (3.9) and the determinant of H are both positive for the estimated least-squares coefficients, than A and B minimize $S(A,B)$. The first condition is satisfied. For the second condition we start with computing the determinant.

$$\begin{aligned} det(H) &= \frac{\partial^2 S(A,B)}{\partial A^2}\cdot\frac{\partial^2 S(A,B)}{\partial B^2} - \Big(\frac{\partial^2 S(A,B)}{\partial A\,\partial B}\Big)^2 \\ &= 2n\cdot\sum_{i=1}^{n} 2X_i^2 - \Big(\sum_{i=1}^{n} 2X_i\Big)^2 \\ &= 4n\sum_{i=1}^{n} X_i^2 - 4\Big(\sum_{i=1}^{n} X_i\Big)^2 \end{aligned} \tag{3.12}$$

Now we use the Cauchy-Schwartz inequality (from [3]):

$$|<u,v>|^2 \le <u,u><v,v>, \tag{3.13}$$

where u and v are vectors and $<.,.>$ denotes an inner product.

Writing (3.13) further out, we get

$$\Big(\sum_{i=1}^{n} u_i v_i\Big)^2 \le \sum_{i=1}^{n} u_i^2\cdot\sum_{i=1}^{n} v_i^2 \tag{3.14}$$

Taking v as an vector where all entries are equal to one, we get from (3.14)

$$\Big(\sum_{i=1}^{n} u_i\Big)^2 \le n\cdot\sum_{i=1}^{n} u_i^2 \tag{3.15}$$

By using (3.15), we can conclude that (3.12) is positive regardless of the value of X_i. So A and B minimize $S(A,B)$.

The equations that we obtain by setting the partial derivatives equal to zero are called the least-squares normal equations. When the value of the explanatory variable X is zero, the intercept A is the fitted value of the response variable Y. The coefficient B represents the average change in Y with an increase of one unit in X. We estimate α and β by the least-squares coefficients.

The statistical model for multiple regression with k explanatory variables and n observations with $1 \le i \le n$ in matrix form is

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & \\ \vdots & & \ddots & \\ 1 & X_{n1} & \dots & X_{nk} \end{bmatrix}\cdot\begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \tag{3.16}$$

where the regression equation for the i-th element of the response variable is

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i. \tag{3.17}$$

We can write (3.16) also as $Y = \mathbf{X}\beta + \epsilon$, where $\mathbf{X}$ is called the design matrix. The least-squares coefficient vector $\hat{\beta}$ for the multiple regression model can be deduced in a same manner as in a simple regression model. The residual vector is

$$(Y - \hat{Y})^2 = (Y - \mathbf{X}\hat{\beta})^2 \tag{3.18}$$

Compute the partial derivative with respect to $\hat{\beta}$.

$$\begin{aligned}
\frac{\partial(Y-\mathbf{X}\hat{\beta})^2}{\partial\hat{\beta}} &= \frac{\partial}{\partial\hat{\beta}}\Big((Y-\mathbf{X}\hat{\beta})^T(Y-\mathbf{X}\hat{\beta})\Big) \\
&= \frac{\partial}{\partial\hat{\beta}}\Big(Y^TY - Y^T\mathbf{X}\hat{\beta} - \hat{\beta}^T\mathbf{X}^TY + \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta}\Big) \\
&= \frac{\partial}{\partial\hat{\beta}}\Big(Y^TY - (Y^T\mathbf{X}\hat{\beta})^T - \hat{\beta}^T\mathbf{X}^TY + \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta}\Big) \\
&= \frac{\partial}{\partial\hat{\beta}}\Big(Y^TY - \hat{\beta}^T\mathbf{X}^TY - \hat{\beta}^T\mathbf{X}^TY + \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta}\Big) \\
&= 0 - \mathbf{X}^TY - \mathbf{X}^TY + (\mathbf{X}^T\mathbf{X}\hat{\beta} + \mathbf{X}^T\mathbf{X}\hat{\beta}) \\
&= -2\mathbf{X}^TY + 2(\mathbf{X}^T\mathbf{X}\hat{\beta}) \\
&= -2\mathbf{X}^T(Y-\mathbf{X}\hat{\beta}) \qquad (3.19)
\end{aligned}$$

Set (3.19) equal to zero and compute $\hat{\beta}$:

$$\begin{aligned}
& -2\mathbf{X}^T(Y-\mathbf{X}\hat{\beta}) = 0 \\
\Longrightarrow & -2\mathbf{X}^TY + 2\mathbf{X}^T\mathbf{X}\hat{\beta} = 0 \\
\Longrightarrow & \boldsymbol{X}^{\boldsymbol{T}}\mathbf{X}\hat{\beta} = \boldsymbol{X}^{\boldsymbol{T}}Y \\
\Longrightarrow & \hat{\beta} = (\boldsymbol{X}^{\boldsymbol{T}}\mathbf{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}Y \qquad (3.20)
\end{aligned}$$

In the last step we have assumed that $\boldsymbol{X}$ has full rank. This implies that $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is invertible. We will show this through an indirect proof.
If $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is not invertible, its columns are linear dependent. Then there exists a vector $p \neq 0$ such that $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}p = 0$. Then we can write

$$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}p = 0 \Longrightarrow p^T\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}p = p^T0 \Longrightarrow (\boldsymbol{X}p)^T(\boldsymbol{X}p) = 0 \Longrightarrow \boldsymbol{X}p = 0 \qquad (3.21)$$

Since $p \neq 0$, this means $\boldsymbol{X}$ has linear dependent columns. Therefore, it is not invertible and thus has not full rank.

So the fitted values are given by

$$\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\boldsymbol{X}^{\boldsymbol{T}}\mathbf{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}Y. \qquad (3.22)$$

3.2. Model assumptions

The model assumptions for simple linear regression rely on the behaviour of the error term ϵ_i. Here follows a list of several assumptions:

- **Linearity**: The regression model is linear in its parameters. Furthermore $E[\epsilon_i] = 0$.
- **Constant variance**: Also known as homoscedasticity. $Var[\epsilon_i] = \sigma_\epsilon^2$ for all values of the explanatory variable.
- **Normality**: The residuals are normally distributed: $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$.
- **Independence**: For $i \neq j$, any pair of residuals ϵ_i and ϵ_j are independent.
- Fixed X or, if random, measured without error and independent of ϵ_i.
- X is not **invariant**. The values of the explanatory variable cannot all be the same. It is not possible to fit a line to data in which the explanatory variable is invariant.

We assume that the data we use is measured without error. The assumptions for a multiple regression model are identical to the assumptions for simple linear regression. Furthermore, we assume that no explanatory variable is a perfect linear function of the others. If this is the case, we have collinearity which is not desirable for our model. More about collinearity can be found in the next chapter.

3.3. Dummy variables

Up to now, we only discussed regression models with quantitative variables. However, as we have seen in the previous chapter not all our data consists of quantitative variables. For example, we have energylabels and districts. To incorporate these qualitative variables into our regression model, we can use dummy variable regressors.

Qualitative and/or categorical explanatory variables are also called factors. Factors can be dichotomous or polytomous. The former consists of two categories, the latter of more than two. Almost all of our qualitative variables are polytomous factors. We start by explaining both types of factors.

3.3.1. Dichotomous factors

Let D be a dummy-variable regressor for a factor of two categories. Then we can incorporate D for observation i in our (simple) regression equation by writing

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_i + \epsilon_i, \tag{3.23}$$

where

$$D_i = \begin{cases} 0 & \text{for category 1 ,} \\ 1 & \text{for category 2} \end{cases}$$

Suppose that the first category in D means a house is not located in the municipality Delft and the second category the opposite. Let Y be the value of a house and X the size of the living space in m^2. Then $D_i = 1$ in equation (3.23), when a house is located in Delft. If a house is outside Delft, $D_i = 0$ and the term $\beta_2 \cdot D_i$ is equal to zero. This omitted category serves as a baseline to which the other category is compared. We call category 1 the baseline (or reference) category.

3.3.2. Polytomous factors

Let Y and X be again the value of a house and the size of the living space, respectively. Suppose that we only consider houses with energylabel A, B and C. To incorporate a three-category factor in our regression equation, we need two dummy regressors D_1 and D_2.

A regression equation with a polytomous factor can be written as

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot D_{i1} + \beta_4 \cdot D_{i2} + \epsilon_i, \tag{3.24}$$

where the values of D_1 and D_2 can be found in the following table.

Energylabel	D_1	D_2
A	1	0
B	0	1
C	0	0

Here, energylabel C serves as the baseline category. We used only three different energylabels, but in practice we have seven different energylabels (A to G). In general, for a factor with n categories, we need $n-1$ dummy regressors.

In this research, we use factors with many categories. For instance, the number of districts or neighborhoods in a municipality. Delft has 13 district, but currently more than 80 neighborhoods.

3.4. Interaction terms

Two explanatory variables X_1 and X_2 interact when the effect on the response variable of X_1 depends on the value of X_2 (or vice versa), whether or not they are statistically related. Therefore, interaction should not be confused with correlation. Interaction does not refer to the relationship between the explanatory variables, but to the manner in which they combine to affect a response variable. Interaction regressors can be incorporated in a regression model by taking products of dummy regressors with quantitative explanatory variables. The principle of marginalty states that a regression model including an interaction term, should also include the low-order relatives of that term. Otherwise, we are not able to interpret the main effects of explanatory variables that interact.

Let us show a regression model with an interaction term.

$$Y_i = \beta_1 \cdot X_{i1} + \beta_2 \cdot D_{i2} + \beta_3 \cdot X_{i1} \cdot D_{i2} \tag{3.25}$$

Suppose that X_1 is the size of the living space in m^2 and that D_2 is a dichotomous factor for two provinces, where

$$D_2 = \begin{cases} 0 & \text{if property is located in Noord-Holland} \\ 1 & \text{if property is located in Zuid-Holland} \end{cases}.$$

Now $X_{i1} \cdot D_{i2}$ is the interaction regressor. β_3 is the corresponding regressor coefficient.

3.5. Regression diagnostics

After running a regression model, we measure the accuracy of the model. We use various methods and check whether the model assumptions are met.

3.5.1. Multiple correlation coefficient

In the first paragraph of this chapter we introduced the sum of squared residuals. In multiple regression models we also quite often deal with other types of sums of squares, namely the total sum of squares (TSS) and the regression sum of squares (RegSS).

$$\text{TSS} = \sum_i (Y_i - \bar{Y})^2 \tag{3.26}$$

$$\text{RegSS} = \sum_i (\hat{Y}_i - \bar{Y})^2 \tag{3.27}$$

Note that

$$\text{RegSS} = \text{TSS} - \text{RSS}, \tag{3.28}$$

where RSS is the residual sum of squares. The RegSS tells us how much of the variation in the dependent variable the model explains. The ratio of RegSS to TSS represents the proportion of variation in the response variable captured by the regression model and is defined as R^2, the squared multiple correlation.

$$R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} \tag{3.29}$$

By convention, the multiple correlation coefficient is the positive square root of R^2. We can also interpret the multiple correlation as the simple correlation between the fitted values and the observed values. Therefore $0 \leq R^2 \leq 1$. In multiple regression settings, the R-squared will never decrease as more variables are included in the model. That is why the adjusted R^2 is the preferred measure as it adjusts for the number of variables considered. The formula for the adjusted R^2, which we write as $\tilde{R}^2$, is:

$$\begin{aligned}\tilde{R^2} &= 1 - \frac{(1-R^2)\cdot(n-1)}{(n-k-1)} \\ &= 1 - \frac{(\frac{\text{TSS}-\text{RegSS}}{\text{TSS}})\cdot(n-1)}{(n-k-1)} \\ &= 1 - \frac{(\frac{\text{RSS}}{\text{TSS}})\cdot(n-1)}{(n-k-1)} \\ &= 1 - \Big[\frac{\text{RSS}}{n-k-1}\cdot\frac{n-1}{\text{TSS}}\Big],\end{aligned} \tag{3.30}$$

where n is the sample size and k the number of regressors. So $n-k-1$ is the degrees of freedom. $\tilde{R^2}$ can take negative values as well. A value close to zero indicates that the regression model explains almost nothing about the variation in the response variable. If $\tilde{R^2}$ is close to 1, much of the variation in the response variable is explained by the regression model.

3.5.2. Influential data

The regression models for our problem that we introduce in the next chapter contain many variables. Some of them have more influence on the response variable than others. An added-variable plot is an useful influence graph which gives us insight in the joint influence of the predictors on the regression coefficients.
Consider a multiple regression model where the first predictor is omitted and let $\epsilon_{Y_i}^{(1)}$ be the corresponding residual.

$$Y_i^{(1)} = \alpha_{Y_i^{(1)}} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_{Y_i^{(1)}} \tag{3.31}$$

The (1) in $Y_i^{(1)}$ denotes that the first predictor is omitted. We can also define $\epsilon_{X_i^{(1)}}$ as the residual of the regression of X_1 on the other predictor variables:

$$X_i^{(1)} = \alpha_{X_i^{(1)}} + \gamma_2 X_{i2} + \cdots + \gamma_k X_{ik} + \epsilon_{X_i^{(1)}} \tag{3.32}$$

For each variable j, with $j = 1, \dots k$, an added-variable plot is constructed by plotting the residuals $\epsilon_{Y^{(j)}}$ versus $\epsilon_{X^{(j)}}$ obtained by the least-squares method. The slope of the least-squares regression of $\epsilon_{Y^{(j)}}$ on $\epsilon_{X^{(j)}}$ is equal to the least-squares slope β_j in the multiple regression model without omitted variables.

3.5.3. Studentized residuals

Outliers can cause problems in linear models: they can influence the results or indicate that the model does not perform well, even if the opposite is true. Most regression diagnostics and model assumptions are based on the residuals. Outliers will have large residuals (in absolute value). These large residuals give a wrong representation of the model. Instead of looking at the residuals, we can look at the studentized residuals.
Let E_i be the residual for observation i. The studentized residual for i is defined as

$$E_i^* = \frac{E_i}{S_{E(-i)}\cdot\sqrt{1-h_i}}, \tag{3.33}$$

where $S_{E(-i)}$ is the standard error of the regression on the model where the i-th observation is deleted. In general,

$$S_E = \sqrt{\frac{\sum E_i^2}{(n-k-1)}}, \tag{3.34}$$

where n is the sample size and k the number of regressors. h_i is the hat-value and is a measure of leverage. Leverage measures the distance between values of an independent variable for different observations. Data points with high leverage are observations with an outlying X value. In simple linear regression the hat-values measure the distance from the mean of X:

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^n (X_j - \overline{X})^2} \tag{3.35}$$

Equation (3.22) can be written as $\hat{Y} = H \cdot Y$, where $\mathbf{H} = \mathbf{X}(\boldsymbol{X}^T\mathbf{X})^{-1}\boldsymbol{X}^T$. H is called the hat-matrix. In a multiple regression model, the hat-values can be obtained by the hat-matrix. In that case the hat-value is defined as $h_i \equiv h_{ii}$. E_i^* is less sensitive to outliers than E_i and therefore useful for further analysis.

3.5.4. Collinearity

When there is a perfect linear relationship between two explanatory variables, we say that they are perfectly collinear. In mathematical terms, X_1 and X_2 are perfectly collinear if there exists constants c_1 and c_2, where they are not both equal to zero, such that $X_{i2} = c_1 \cdot X_{i1} + c_2$. In that case the correlation between X_1 and X_2 is 1 or -1.
We speak of multicollinearity, when for k explanatory variables we have

$$c_1 \cdot X_{i1} + c_2 \cdot X_{i2} + \cdots + c_k \cdot X_{ik} = c_0 \tag{3.36}$$

where the constant $c_1, \ldots c_k$ are not all 0. c_0 is a constant as well. In practice, we barely have perfect (multi)collinearity. Usually, when there is collinearity, it is less than perfect, but still strong enough to pay attention to.

When we have (multi)collinearity, the least-squares normal equations do not have a unique solution [1]. Furthermore, small changes (like adding an extra explanatory variable) in the model can cause large changes in the regression coefficients. Collinearity increases the sampling variances of the least-squares estimators. This will affect the values of the individual predictions. Therefore, it is important to keep track of collinearity in the model.

Due to the simplicity and direct interpretation, the variance-inflation factor (VIF) is the basic diagnostic for collinearity. The variance-inflation factor is defined as

$$VIF_j = \frac{1}{1 - R_j^2} \tag{3.37}$$

where R_j^2 is the the squared multiple correlation for the regression of X_j on the other explanatory variables. VIF_j indicates the deletorious impact of collinearity on the least-squares estimate of the regression coefficient for variable j. When R_j^2 approaches 0.9, the precision of the estimates of the coefficients is seriously degraded. This corresponds to a VIF around 5. Therefore, we will use this value when examining collinearity in our model.

The VIF cannot be applied to models that include dummy variables and polynomial regressors, since we cannot obtain the squared multiple correlation coefficient for a regression model with a qualitative response variable. As alternative we can use the generalized variance-inflation factor (GVIF).

Consider a linear regression model with categorical variables. Let us write the linear model as

$$\underset{(n\times 1)}{\mathrm{Y}} = \alpha \times \underset{(n\times 1)}{1} + \underset{(n\times p)}{X_1} \times \underset{(p\times 1)}{\beta_1} + \underset{(n\times(k-p))}{X_2} \times \underset{((k-p)\times 1)}{\beta_2} + \underset{(n\times 1)}{\varepsilon}. \tag{3.38}$$

The matrix X_1 contains p dummy regressors of interest and the remaining variables are in X_2. α is the intercept. β_1 contains the coefficients for $X_1$and β_2 contains the coefficients for X_2. ε is the vector with errors. Let R_{11} and R_{22} be the correlation matrix for X_1 and X_2, respectively. Furthermore, let R be the correlation matrix among all variables. The GVIF for explanatory variable 1 is defined as

$$GVIF_1 = \frac{det(R_{11}) \cdot det(R_{22})}{det(R)} \tag{3.39}$$

To make generalized variance-inflation factors comparable across dimensions, we can take $GVIF^{\frac{1}{2\cdot df}}$, where df is the number of regressors for each explanatory variable instead of $GVIF$ [5].

3.5.5. Checking the model assumptions

In paragraph 2 of this chapter we listed the assumptions that are made for linear regression. The main assumptions that we have to check are normality, linearity and constant error variance. The other assumptions are usually already accounted for in the data.

Normality

One of the main assumptions in linear regression is that the residuals are normally distributed. The Gauss-Markov theorem states that the least-squares estimator is the most efficient linear unbiased estimator, when the errors have zero mean and are uncorrelated. However, for other type of distributions than the normal distribution the efficiency decreases particularly. An example is a heavy tailed distribution. To check whether the residuals are normally distributed we can use a QQ-plot, that is a quantile comparison plot. It is a graphical method to examine the distribution of the residuals. A QQ-plot compares the quantiles of two probability distributions. In our case, we plot the quantiles of the empirical sample distribution of the (studentized) residuals, against quantiles of the (standard) normal distribution. If the points in the QQ-plot lie more less on the line $y = x$, the two compared distributions are approximately similar. If for instance the point are too far from $y = x$ in the tails, we have most likely a heavy tailed distribution. This means that we have (too many) outliers. Consequently we should look at other estimators than least-squares. Robust estimators can be helpful in case of heavy tailed distributions. If it turns out that we need to use robust estimators, we will discuss this later.
As alternative for a QQ-plot, we could also use a histogram with many bars or a density plot to examine the distribution of the (studentized) residuals. This will display the tail behaviour and skewness.

Linearity

In linear regression we make the assumption that there is a linear relationship between the explanatory variables and the response variable. Non-linear patterns can be detected through a plot of the (studentized) residuals versus the fitted values. If we have in such a plot equally spread of the residuals around a horizontal line, without finding any distinctive pattern, we likely have no non-linear patterns. However, such a scatterplot considers all explanatory variables together, which is not the optimal way to detect non-linearity since it can be misleading and hide patterns. Better is to look at the relationship between the response variable and an explanatory variable, taking into account the other explanatory variables as well.
Define the partial residual for explanatory variable j and observation i as

$$E_i^{(j)} = E_i + B_j \cdot X_{ij}. \tag{3.40}$$

A component-plus-residual plot is constructed by plotting $E^{(j)}$ against X_j. So B_j is the slope of the simple linear regression of $E^{(j)}$ on X_j. A component-plus-residual plot can reveal non-linear patterns of a variable when the other variables are involved in the model, even tough it is constructed by plotting a simple linear regression equation. Once non-linearity is detected, we might be able to apply a variable transformation to the explanatory variable to make it linear.

Constant error variance

Homoscedasticity (or constant error variance) is a main assumption in linear regression. A non-constant variance for the residuals has a consequence that the efficiency of the least-squares estimators decreases. Therefore, we should always check whether we have homoscedasticity or heteroscedasticity (non-constant error variance).

Just as with the previous two assumptions, we have for the error variance a graphical examination as well. In particular, we can use a plot of the absolute values of the studentized residuals versus the fitted values. If in this plot we have a equally spread of the residuals around a horizontal line, we do not have heteroscedasticity. If we detect a certain pattern, we probably have heteroscedasticity. For example, it is common for the variance of the residuals to increase when the value of the response variable increases. A transformation of the response variable can stabilize the variance.

3.5.6. Non-linear regression

Up till now we only discussed regression models that are linear in the regressors. However, we can extend these models for example by including polynomial regressors that generate several regressors for one explanatory variable. We also have the option to transform explanatory variables or the response variable. These modifications are mostly necessary to met the regression assumptions or to have a better coherence between the data and model.

We can apply a function $f(\cdot)$ to a quantitative explanatory variable. This function can be for example a polynomial, which gives rise to multiple regressors (linear, quadratic, cubic etc.), or a logarithmic function. When we have multiple regressors, each one gets its own coefficient. We assume that the function $f(\cdot)$ does not contain unknown parameters. These types of non-linear regression models are linear in parameters and can be fitted by the linear least-squares method that we have discussed in the previous paragraphs.

3.5.7. Data transformation

We already mentioned the possibility to transform the data. For instance, in Subsection 3.5.5 we already mentioned that a variable transform can help to make non-linear relationships more linear. In this paragraph we describe the guidelines we follow to decide whether we need to transform the data.

Many options exists to transform the data. A useful transformation is the group of the family of power and roots where we transform X to X^p. For $p < 0$ we get a inverse power transformation. A special type of power transform is the Box-Cox transformation that is defined as

$$X \mapsto X^{(p)} \equiv \frac{X^p - 1}{p}. \tag{3.41}$$

For $p = 0$ we apply the logarithmic transformation, where X^p and $X^{(p)}$ become $log(X)$. Here we use the natural logarithm. The difference between X^p and $X^{(p)}$ as defined above is that $X^{(p)}$ is a linear function of X^p and is more transparent in revealing the essential unity of the family of powers and roots.

To decide how we should take p, we can use Tukey and Mosteller's bulging rule.

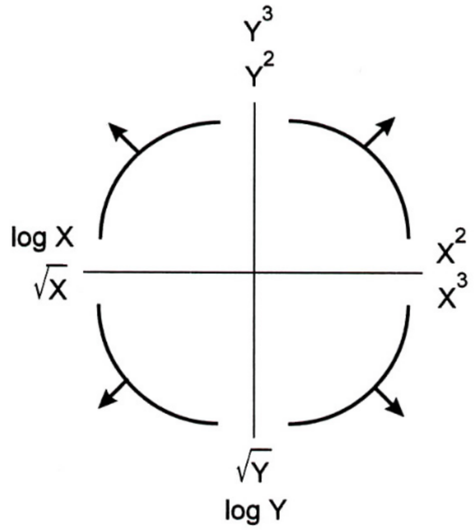


Figure 3.1: Tukey and Mosteller's bulging rule: The direction of the bulge indicates the direction of the power transformation of Y and/or X to straighten the relationship between them (from [24]).

Depending on the direction of the bulge of the data we decide to go up or down the ladder of power transforms. At the end, it depends on the data how much we go up (e.g. X^2, X^3) or down (e.g. $X^{1/2}$ or $X^{1/3}$) and what fits best to the model.

The log-transform is a broadly used transformation. It makes the differences between large values smaller. This can help when we have outliers. Moreover, it can help to make positive skewed distributions more symmetric. This makes the examination of the data more convenient. On the other hand, for a negative skewed distribution we can use a transformation ascending the ladder of powers.

In particular, data transformations are most of the time necessary when we have a non-linear relationship among variables. It is important to study the related scatterplots, added-variable plots and component-plus-residual-plots to get a good impression of the cohesion and influence of the different variables, before we start with a possible data transformation.

3.6. Robust estimators

As mentioned earlier, the efficiency of the least-squares approach is vitiated by heavy tailed errors and many outlying values. Robust regression is an alternative to deal with heavy tailed error distributions. Robust estimators are resistant to outliers because they down-weight them. Furthermore, they are almost as efficient as least-squares when the error distribution is normal and much more efficient when the errors are heavy tailed.

Before we start with robust estimators for a multiple regression model, we will first discuss robust estimators in a much simpler setting. We consider the linear model

$$Y_i = \mu + \epsilon_i \tag{3.42}$$

and want to estimate μ which is the center of some symmetric distribution, called the location. ϵ_i is the error term. Let the fitted value of (3.42) be given by $\hat{Y}_i = \hat{\mu}$. Then the residual is given by $E = Y - \hat{Y} = Y - \hat{\mu}$. We would like to minimize the objective function

$$\sum_{i=1}^{n} \rho(E_i). \tag{3.43}$$

For the least-squares estimator we have $\rho_{LS}(E_i) = (Y_i - \hat{\mu})^2$. To assess the influence of outlying values on the objective function, we can look at the influence function $\psi(E)$ of the estimator which has the same shape as the derivative of the objective function, $\psi(E) \equiv \rho'(E)$. For the least-squares estimator we have

$$\psi_{LS}(E) \equiv \rho'_{LS}(E) = 2E. \tag{3.44}$$

Note that $\psi_{LS}(E)$ is not bounded and therefore will be far form zero for outyling values. As consequence, we have to look for other estimators.

An option is to look at the least absolute values (LAV), also known as least absolute deviations (LAD), to minimize the sum of residuals. We have $\rho_{LAV}(E_i) = |Y_i - \hat{\mu}|$. The derivative of $\rho_{LAV}(E)$ gives the shape of the influence function.

$$\psi_{LAV}(E) \equiv \rho'_{LAV}(E) = \begin{cases} 1 & \text{for } E > 0 \\ 0 & \text{for } E = 0 \\ -1 & \text{for } E < 0 \end{cases}. \tag{3.45}$$

Stricly speaking, the derivative of $\rho_{LAV}(E)$ is not defined at $E = 0$. By convention we set $\rho'_{LAV}(0) \equiv 0$. As we can see, $\psi_{LAV}(E)$ is bounded and therefore much more resistant to outliers.
Estimators that can be expressed as minimizing an objective function $\sum_{i=1}^{n} \rho(E_i)$ are called M estimators. Two common choices of objective functions are the Huber and Tukey's biweight (or bisquare) functions.

In the next table the objective function and influence functions can be found for the Huber and Tukey's biweight estimator, together with the earlier discussed estimators.

Table 3.1: Overview of the M-estimators and their corresponding objective functions and influence functions.

Estimator	Objective function	Influence function
Least-squares	$\rho_{LS}(E) = (E)^2$	$\psi_{LS}(E) \equiv 2E$
LAD	$\rho_{LAV}(E) = \lvert E\rvert$	$\psi_{LAV}(E) \equiv \begin{cases} 1 & \text{for } E > 0 \\ 0 & \text{for } E = 0 \\ -1 & \text{for } E < 0 \end{cases}$
Huber	$\rho_H(E) = \begin{cases} \frac{1}{2}E^2 & \text{for } \lvert E\rvert \leq k \\ k\lvert E\rvert - \frac{1}{2}k^2 & \text{for } \lvert E\rvert > k \end{cases}$	$\psi_H(E) \equiv \begin{cases} k & \text{for } E > k \\ E & \text{for } \lvert E\rvert \leq k \\ -k & \text{for } E < -k \end{cases}$
Tukey's biweight	$\rho_{BW}(E) = \begin{cases} \frac{k^2}{6}\left[1-\left(1-\left(\frac{E}{k}\right)^2\right)^3\right] & \text{for } \lvert E\rvert \leq k \\ \frac{k^2}{6} & \text{for } \lvert E\rvert > k \end{cases}$	$\psi_{BW}(E) \equiv \begin{cases} E\left[1-\left(\frac{E}{k}\right)^2\right]^2 & \text{for } \lvert E\rvert \leq k \\ 0 & \text{for } \lvert E\rvert > k \end{cases}$

The value k in the table for the Huber and biweight estimator is called the tuning constant. It is defined as $k = c \cdot S$ for some constant c. S is the scale and defines the spread of the variable Y. A measure of scale that is common and robust is the median absolute deviation (MAD) which is defined as

$$\text{MAD} \equiv \text{median}\left|Y_i - \text{median}(\hat{\mu})\right|. \tag{3.46}$$

M-estimators are a generalization of maximum likelihood estimators. The latter attain the Cramér-Rao variance asymptotically. This is called asymptotic efficiency. For $k = 1.345S$, the Huber estimator has 95% efficiency when the population of Y is normal. For the bisquare estimator the same holds for $k = 4.685S$. This means that the size of the asymptotic variance of a sample from the normal distribution is 95% of the size of the asymptotic variance of the M-estimator.

Remember that our goal is to estimate μ in (3.42). Calculation of M-estimators can be done via an iterative procedure. We will use the iterative weighted least squares (IWLS) method which we describe soon.

An estimation $\hat{\mu}$ can be obtained by differentiating the objective function with respect to $\hat{\mu}$ and set it equal to zero. In mathematical terms:

$$\begin{aligned} &\sum_{i=1}^{n} \rho'(E_i) = 0 \\ \Longrightarrow &\sum_{i=1}^{n} \rho'(Y_i - \hat{\mu}) = 0 \\ \Longrightarrow &\sum_{i=1}^{n} \psi(Y_i - \hat{\mu}) = 0. \end{aligned} \tag{3.47}$$

Define the weight function as

$$w(E) = \frac{\psi(E)}{E}, \tag{3.48}$$

and set

$$w_i \equiv w(Y_i - \hat{\mu}). \tag{3.49}$$

Then we can rewrite (3.47):

$$
\begin{aligned}
&\sum_{i=1}^{n} \psi(Y_i - \hat{\mu}) = 0 \\
\Longrightarrow &\sum_{i=1}^{n} (Y_i - \hat{\mu}) \cdot \frac{\psi(Y_i - \hat{\mu})}{(Y_i - \hat{\mu})} = 0 \\
\Longrightarrow &\sum_{i=1}^{n} (Y_i - \hat{\mu}) \cdot w(Y_i - \hat{\mu}) = 0 \\
\Longrightarrow &\sum_{i=1}^{n} (Y_i - \hat{\mu}) \cdot w_i = 0
\end{aligned}
\tag{3.50}
$$

From (3.50) we can get a solution for $\hat{\mu}$:

$$
\begin{aligned}
&\sum_{i=1}^{n} (Y_i - \hat{\mu}) \cdot w_i = 0 \\
\Longrightarrow &\sum_{i=1}^{n} (Y_i \cdot w_i - \hat{\mu} \cdot w_i) = 0 \\
\Longrightarrow &\sum_{i=1}^{n} Y_i \cdot w_i = \sum_{i=1}^{n} \hat{\mu} \cdot w_i \\
\Longrightarrow &\hat{\mu} = \frac{\sum_{i=1}^{n} Y_i \cdot w_i}{\sum_{i=1}^{n} w_i}
\end{aligned}
\tag{3.51}
$$

Now since $\hat{\mu}$ depends on the weights w_i and vice versa, the iterative weighted least squares method requires that we start with an initial estimate $\hat{\mu}^{(0)}$, compute initial weights and update the value of $\hat{\mu}$ until it converges.

Algorithm 1 Calculate $\hat{\mu}$ with IWLS

Choose: $\hat{\mu}^{(0)}$ and c

Calculate $S^{(0)}$ and use this to find k
$w_i^{(0)} = w(Y_i - \hat{\mu}^{(0)})$
$l = 1$
$\hat{\mu}^{(l)} = \dfrac{\sum_{i=1}^{n} Y_i \cdot w_i^{(l-1)}}{\sum_{i=1}^{n} w_i^{(l-1)}}$
while $|\hat{\mu}^{(l)} - \hat{\mu}^{(l-1)}| \not\approx 0$ **do**
    $l = l + 1$
    Calculate $S^{(l)}$ and k
    $w_i^{(l-1)} = w(Y_i - \hat{\mu}^{(l-1)})$
    $\hat{\mu}^{(l)} = \dfrac{\sum_{i=1}^{n} Y_i \cdot w_i^{(l-1)}}{\sum_{i=1}^{n} w_i^{(l-1)}}$
end while

$\hat{\mu} = \hat{\mu}^{(l)}$

Note that the scale is required to calculate the tuning constant, which is necessary to compute the weights.

In the table below the weight functions can be found for the earlier discussed M-estimators.

Table 3.2: Overview of the M-estimators and their corresponding weight function.

Estimator	Weight function
Least-squares	$w_{LS}(E) = 1$
LAD	$w_{LAV}(E) = \frac{1}{\lvert E\rvert}$ for $E \neq 0$
Huber	$w_H(E) = \begin{cases} 1 & \text{for } \lvert E\rvert \leq k \\ \frac{k}{\lvert E\rvert} & \text{for } \lvert E\rvert > k \end{cases}$
Bisquare	$w_{BW}(E) = \begin{cases} \left[1 - \left(\frac{E}{k}\right)^2\right]^2 & \text{for } \lvert E\rvert \leq k \\ 0 & \text{for } \lvert E\rvert > k \end{cases}$

In Figure 3.2 and 3.3 we can see the objective, influence and weight function for the Huber and bisquare estimator. The tuning constant is set to $k = 1$ to make both graphs comparable. As we can see, both estimators are much more resistant to outliers than the least-squares estimator in terms of influence and weight.

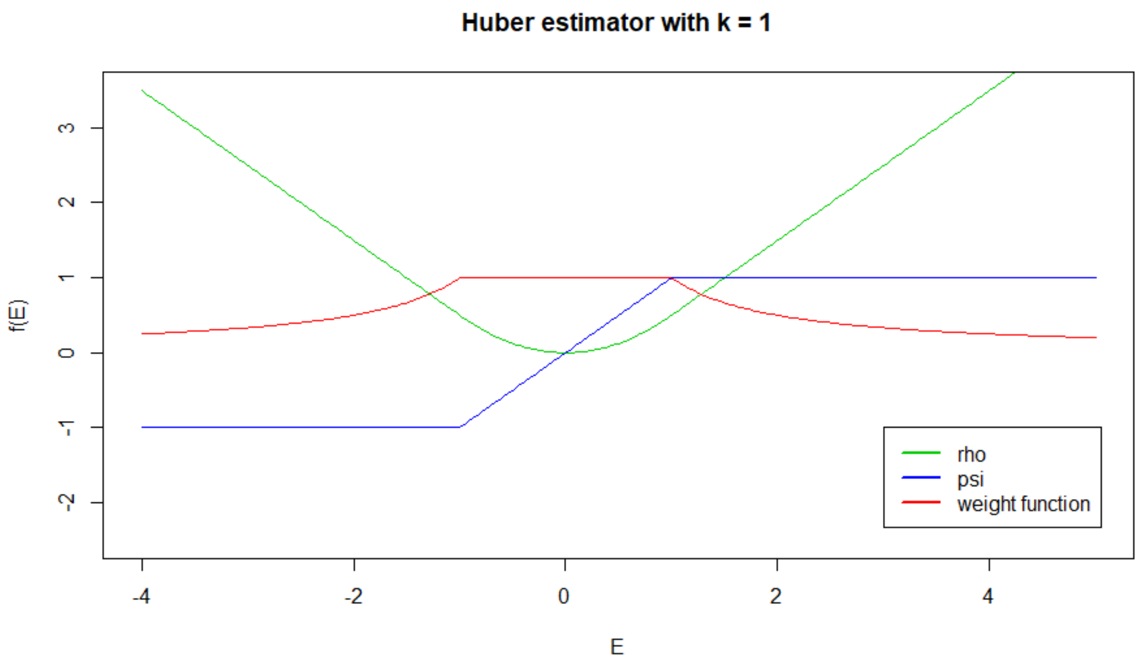


Figure 3.2: Objective, influence and weight functions for the Huber estimator

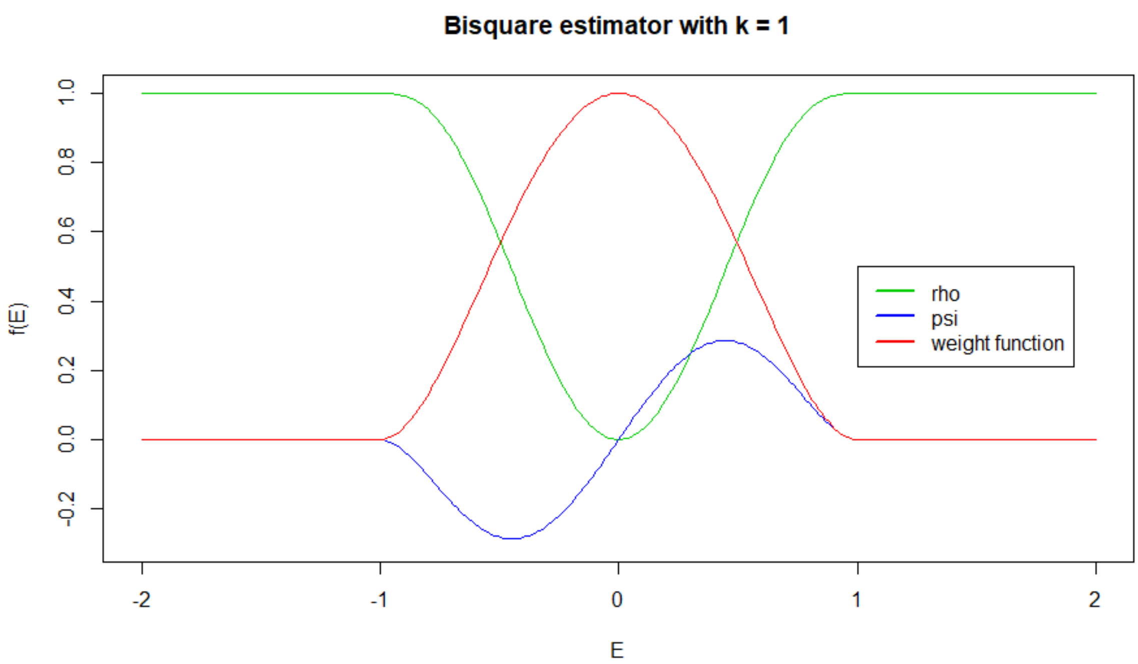


Figure 3.3: Objective, influence and weight functions for the bisquare estimator

3.7. Robust regression

In the previous section we discussed robust estimators applied to the simple model (3.42). In this section we discuss robust estimators applied to a multiple regression model. We consider the model as in (3.17)

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i.$$

The fitted value is given by

$$\begin{aligned}\hat{Y}_i &= A + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_k X_{ik} \\ &= \boldsymbol{x}_i^T \cdot \mathbf{b},\end{aligned} \tag{3.52}$$

where the vectors $\mathbf{x}_i$ and $\mathbf{b}$ contain the values of the explanatory variables for observation i and the estimated coefficients, respectively. Now again we want to minimize the objective function of some M-estimator and set its derivative with respect to $\mathbf{b}$ equal to zero. The objective function is

$$\sum_{i=1}^{n} \rho(E_i) = \sum_{i=1}^{n} \rho(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i^T \cdot \mathbf{b}). \tag{3.53}$$

Compute the derivative with respect to $\boldsymbol{b}$ by using the gradient, set it equal to zero and use (3.48) and (3.49) to write

$$\begin{aligned}&\nabla_{\boldsymbol{b}}\Big(\sum_{i=1}^{n} \rho'(E_i)\Big) = 0 \\ \Longrightarrow &\nabla_{\boldsymbol{b}}\Big(\sum_{i=1}^{n} \rho'(Y_i - \mathbf{x}_i^T \cdot \mathbf{b})\Big) = 0 \\ \Longrightarrow &\sum_{i=1}^{n} \psi(Y_i - \mathbf{x}_i^T \cdot \mathbf{b})\mathbf{x}_i = 0 \\ \Longrightarrow &\sum_{i=1}^{n} w_i \cdot (Y_i - \mathbf{x}_i^T \cdot \mathbf{b})\mathbf{x}_i = 0.\end{aligned} \tag{3.54}$$

Now we can find an analogy with the weighted least-squares estimator. Computing the derivative of the objective function of the weighted least-squares estimator and setting this equal to zero, we obtain

$$\begin{aligned}&\nabla_{\boldsymbol{b}}\Big(\sum_{i=1}^{n} w_i E_i^2\Big) = 0 \\ \Longrightarrow &\nabla_{\boldsymbol{b}}\Big(\sum_{i=1}^{n} w_i (Y_i - \mathbf{x}_i^T \cdot \mathbf{b})^2\Big) = 0 \\ \Longrightarrow &2\sum_{i=1}^{n} w_i \cdot (Y_i - \mathbf{x}_i^T \cdot \mathbf{b})\mathbf{x}_i = 0.\end{aligned} \tag{3.55}$$

We can see that solving (3.54) is equivalent to solving (3.55). To find the solution of (3.54), we will use the equation that sets the derivative of the weighted least-squares objective function equal to zero. We will use that:

- $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$
- $\boldsymbol{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$
- $\boldsymbol{b} = \begin{pmatrix} 1 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}$

Let us start with the computation.

$$
\begin{aligned}
&\nabla_{\boldsymbol{b}}\Big(\sum_{i=1}^{n} w_i E_i^2\Big) = 0\\
\Longrightarrow &\nabla_{\boldsymbol{b}}\Big(\sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i^T \cdot \mathbf{b})^2\Big) = 0\\
\Longrightarrow &\nabla_{\boldsymbol{b}}\Big(w_1(y_1 - \mathbf{x}_1^T \cdot \mathbf{b})^2 + \dots + w_n(y_n - \mathbf{x}_n^T \cdot \mathbf{b})^2\Big) = 0\\
\Longrightarrow &\nabla_{\boldsymbol{b}}\Big((y_1 - \mathbf{x}_1^T \cdot \mathbf{b}) \cdot w_1 \cdot (y_1 - \mathbf{x}_1^T \cdot \mathbf{b}) + \dots + (y_n - \mathbf{x}_n^T \cdot \mathbf{b}) \cdot w_n \cdot (y_n - \mathbf{x}_n^T \cdot \mathbf{b})\Big) = 0\\
\Longrightarrow &\nabla_{\boldsymbol{b}}\left(\left[\begin{pmatrix} y_1\\ \vdots\\ \vdots\\ y_n\end{pmatrix} - \begin{pmatrix} \mathbf{x}_1^T\\ \vdots\\ \vdots\\ \mathbf{x}_n^T\end{pmatrix} \cdot \begin{pmatrix} 1\\ b_1\\ \vdots\\ b_k\end{pmatrix}\right]^T \cdot \begin{bmatrix} w_1 & 0 & \dots & 0\\ 0 & w_2 & \dots & \vdots\\ \vdots & 0 & \ddots & 0\\ 0 & \dots & 0 & w_n\end{bmatrix} \cdot \left[\begin{pmatrix} y_1\\ \vdots\\ \vdots\\ y_n\end{pmatrix} - \begin{pmatrix} \mathbf{x}_1^T\\ \vdots\\ \vdots\\ \mathbf{x}_n^T\end{pmatrix} \cdot \begin{pmatrix} 1\\ b_1\\ \vdots\\ b_k\end{pmatrix}\right]\right) = 0\\
\Longrightarrow &\nabla_{\boldsymbol{b}}\left(\left[\begin{pmatrix} y_1\\ \vdots\\ \vdots\\ y_n\end{pmatrix} - \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k}\\ 1 & x_{21} & x_{22} & \dots & x_{2k}\\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk}\end{pmatrix} \cdot \begin{pmatrix} 1\\ b_1\\ \vdots\\ b_k\end{pmatrix}\right]^T \cdot \begin{bmatrix} w_1 & 0 & \dots & 0\\ 0 & w_2 & \dots & \vdots\\ \vdots & 0 & \ddots & 0\\ 0 & \dots & 0 & w_n\end{bmatrix} \cdot \left[\begin{pmatrix} y_1\\ \vdots\\ \vdots\\ y_n\end{pmatrix} - \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k}\\ 1 & x_{21} & x_{22} & \dots & x_{2k}\\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk}\end{pmatrix} \cdot \begin{pmatrix} 1\\ b_1\\ \vdots\\ b_k\end{pmatrix}\right]\right) = 0\\
\Longrightarrow &\nabla_{\boldsymbol{b}}\Big((\mathbf{y} - \mathbf{X}\mathbf{b})^T \cdot \mathbf{W} \cdot (\mathbf{y} - \mathbf{X}\mathbf{b})\Big) = 0,
\end{aligned}
\tag{3.56}
$$

where $\boldsymbol{W}$ is a diagonal matrix with the weights on the diagonal.

We can write this further out.

$$
\begin{aligned}
&\nabla_{\boldsymbol{b}}\Big((\mathbf{y} - \mathbf{X}\mathbf{b})^T \cdot \mathbf{W} \cdot (\mathbf{y} - \mathbf{X}\mathbf{b})\Big) = 0\\
\Longrightarrow &\nabla_{\boldsymbol{b}}\Big(\boldsymbol{y^T W y - y^T W X b - b^T X^T W y + b^T X^T W X b}\Big) = 0\\
\Longrightarrow &\nabla_{\boldsymbol{b}}\Big(\boldsymbol{y^T W y - (y^T W X b)^T - b^T X^T W y + b^T X^T W X b}\Big) = 0\\
\Longrightarrow &\nabla_{\boldsymbol{b}}\Big(\boldsymbol{y^T W y - 2 b^T X^T W y + b^T X^T W X b}\Big) = 0\\
\Longrightarrow &-2\boldsymbol{X^T W y} + \boldsymbol{X^T W X b + (b^T X^T W X)^T} = 0\\
\Longrightarrow &2(\boldsymbol{-X^T W y + X^T W X b}) = 0\\
\Longrightarrow &\boldsymbol{X^T W X b} = \boldsymbol{X^T W y}\\
\Longrightarrow &\boldsymbol{b} = \boldsymbol{(X^T W X)^{-1} \cdot X^T W y}
\end{aligned}
\tag{3.57}
$$

In the last step we have assumed that $(\boldsymbol{X^T W X})$ has full rank, as in the computation of equation (3.20).

To find **b** we can again use the IWLS algorithm. Note that we have now a vicious circle with a dependency among three unknowns. The estimated coefficients depend on the weights, the weights depend on the residuals and the residuals depend on the estimated coefficients. IWLS deals with this by successively updating the quantities.

Algorithm 2 Calculate $\boldsymbol{b}$ with IWLS

Choose: $\boldsymbol{b}^{(0)}$ and c

Calculate $S^{(0)}$ and use this to find k

for i = 1, ..., n **do**
  $E_i^{(0)} = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}^{(0)}$

  $w_i^{(0)} = w(E_i^{(0)})$
end for

$l = 1$
$\boldsymbol{b}^{(l)} = (X^T W X)^{-1}(X^T W y)$, where $W = diag\{w_i^{(l-1)}\}$

while $|\boldsymbol{b}^{(l)} - \boldsymbol{b}^{(l-1)}| \not\approx 0$ **do**
  Calculate $S^{(l)}$ and k

  for i = 1, ..., n **do**
    $E_i^{(l)} = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}^{(l)}$
    $w_i^{(l)} = w(E_i^{(l)})$
  end for

  $l = l + 1$
  $\boldsymbol{b}^{(l)} = (X^T W X)^{-1}(X^T W y)$, where $W = diag\{w_i^{(l-1)}\}$
end while

$\boldsymbol{b} = \boldsymbol{b}^{(l)}$

4
Model for Delft

The Netherlands has approximately 7.5 million dwellings. This is a large number and therefore we take a smaller sample size for our analysis. We start by considering only one municipality: Delft. There is not an underlying reason to chose Delft over others.
We start by exploring the data. We use the variables mentioned in the second chapter and try to make a regression model where we take all regression diagnostics into account.

4.1. Data analysis

Our goal is to set up a regression model for the value of a house. As starting point we use the WOZ value as response variable. The WOZ value gives an indication of the value of a house. In general, a house is most of the time sold for a selling price above the WOZ value. Since we do not have the selling prices available, we have to use the data we have. The very first explanatory variable we consider is the living space. The size of the living space is an important characteristic of a house and plays therefore a key role in the value.

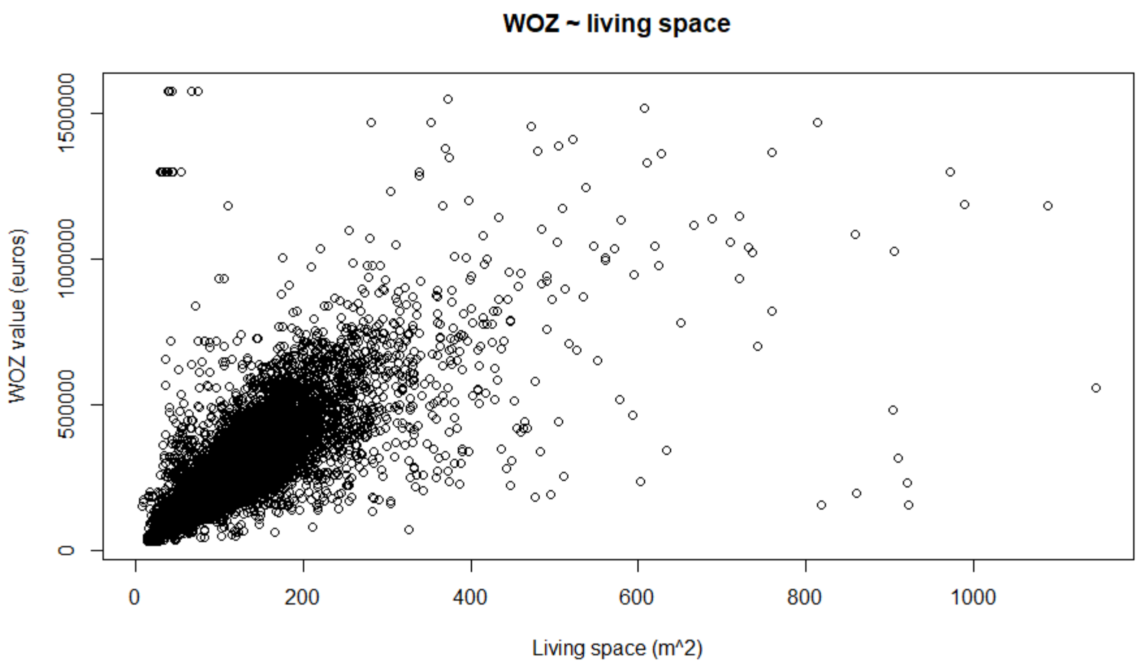


Figure 4.1: WOZ versus living space for properties in Delft

As we can see in Figure 4.1 most of the data points have a living space below 200 m^2 and a WOZ value below 500000 euros. Fitting a straight line through this region will ignore many data points and therefore we would like to stretch out the high dense region. Using Figure 3.1 we can transform one (or both) of the variables down. We try to use a log-transform. It turns out that if we apply a log-transform to one of the variables, we should transform the other variable as well due to a heavy bulge shaped curve. Therefore, we apply a log-transform to both variables. Now we see in Figure 4.2 that there is a better shaped linear relationship between both variables.

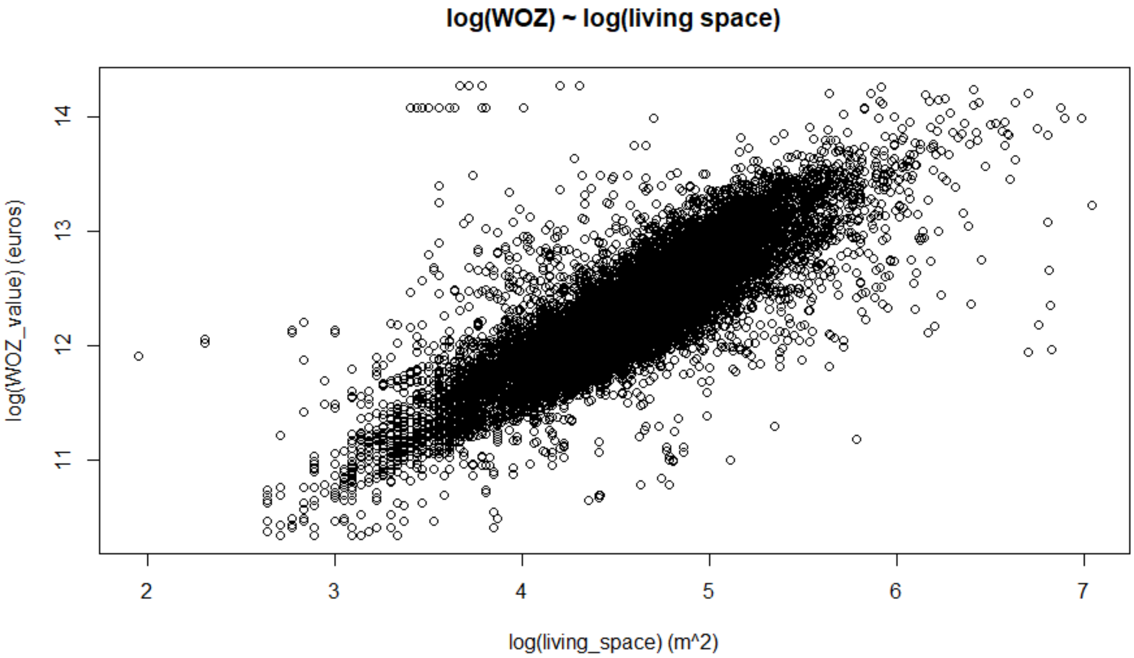


Figure 4.2: A log-transformation for both WOZ and living space for properties in Delft

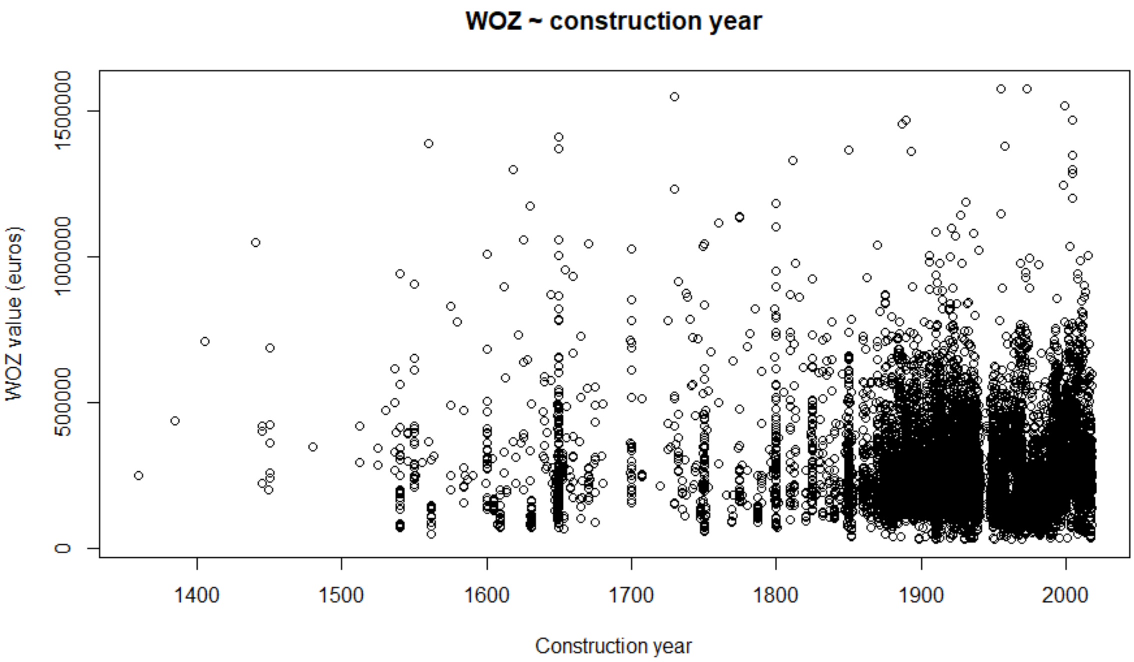


Figure 4.3: WOZ versus construction years for properties in Delft

Another important feature is the construction year. It is clear that the relationship with the WOZ value is not linear, see Figure 4.3. Also we cannot observe a bulge and its direction. Hence, power transformation are ruled out. An option is to split the data into groups and use a categorical variable. When setting up the categories, we strive for groups that do not differ too much in size and that it represents the data well. With the latter we mean for example in this case that houses build in the years after World War II are placed into one group or that a group is formed by houses build in a certain century. Due to the housing shortage after World War II, quantity was more important than quality and therefore most of these houses are not the highest valuated properties.

The next table gives an overview of our grouping of the categories.

Table 4.1: Grouping of the categories for construction year.

Categories for construction years in Delft	
Group name	Time period construction years
17	≤ 1800
18.01	1801-1875
18.76	1876-1900
19.01	1901-1925
19.26	1926-1947
19.48	1948-1974
19.75	1975-2001
20.02	2002-2019

The same strategy we use for all other variables. Plots, transformations and categories for other variables can be found in the Appendix.

We list all the variables and declare whether we use a (power) transformation or dummy variables.

Table 4.2: Variable transformations

Variable transformations	
Variable	**Transformation**
WOZ value	$f(x) = log(x)$
living space	$f(x) = log(x)$
Construction year	factor
Distance to school	$f(x) = x^{1/3}$
Distance to supermarket	factor
Distance to hospital	$f(x) = x^2 + x$
Distance to highway ramp	$f(x) = x^2 + x$
Distance to bus station	factor
Distance to shopping mall	factor
Distance to train station	factor
Distance to residential boulevard	factor
District	factor
Neighborhood	factor
Energylabel	factor
Type of the property	factor
RLBBEV (Habitants dimension for Leefbaarometer)	-
RLBWON (Dwelling dimension for Leefbaarometer)	-
RLBVRZ (Facilities dimension for Leefbaarometer)	-
RLBVEI (Safety dimension for Leefbaarometer)	-
RLBFYS (Physical environment dimension for Leefbaarometer)	-

The five leefbaarometer dimensions, which do not get a transformation, we use on neighborhood level.

4.1.1. Missing data

A problem we often have to deal with is missing data. Not for all properties we have for example the WOZ value or the type. Reasons can be that the WOZ values date from 2017 and that in the years after new properties were build. Or that new neighborhoods are set up and that there is no leefbaarmeter data available, since it is data from 2016. The latter is for us problematic, because two districts disappear from the model. To be able to involve properties with missing data in our model, we can use k-nearest neighbor (kNN) imputation. Imputation is the process of deriving or estimating missing values in a dataset.

In kNN imputation the k nearest neighbors of a property with at least one missing value are determined. The missing value is then imputed by a value computed out of the k nearest neighbors. The nearest neighbors can be found by (a variation of) the Gower distance. The Gower distance between two observations i and j is defined as

$$d_g(i,j) = \frac{\sum_k d_k(i,j) w_{ijk}}{\sum_k w_{ijk}}, \tag{4.1}$$

where k is an index for the explanatory variables, $d_k(i,j)$ is the distance between observation i and j for the explanatory variable with index k and w_{ijk} is the weight.

The distance $d_k(i,j)$ is for categorical variables defined as

$$d_k^{\text{cat}}(i,j) = \begin{cases} 0 & \text{if } X_{ik} = X_{jk} \\ \\ 1 & \text{otherwise} \end{cases} \tag{4.2}$$

and for numerical variables as

$$d_k^{\text{num}}(i,j) = \frac{|X_{ik} - X_{jk}|}{\max\limits_n(X_{nk}) - \min\limits_n(X_{nk})}, \tag{4.3}$$

where n is the number of observations. When $\max_n(X_{nk}) = \min_n(X_{nk})$ in (4.3), $d_k^{\text{num}}(i,j) = 0$.

Remember from the previous chapter that a dichotomous variable only has two categories. Therefore we can also use a binary variable instead of a dichotomous variable. In that case we have two states: 0 and 1. When both states are not equally valuable, so there is a preference on which state should be coded as 0 or 1, we call the binary variable asymmetric. The weight w_{ijk} in (4.1) is defined as

$$w_{ijk} = \begin{cases} 0 & \text{if at least one of } X_{ik}\ ,\ X_{jk} \text{ is missing} \\ 0 & \text{if the variable with index k is asymmetric binary and } X_{ik} = X_{jk} = 0\ . \\ 1 & \text{otherwise} \end{cases} \tag{4.4}$$

Not surprisingly, the k nearest neighbors are the k neighbors with the smallest Gower distance.

After the k nearest neighbors are determined, the median of these neighbors is used as imputation value for the explanatory variable for numerical variables. For categorical variables, the category that most often occurs in the k nearest neighbors is used as imputation. If this results in a draw, a category from these tied category is selected randomly.

We use kNN imputation for missing values for energylabels, property types and leefbaarometer data. A disadvantage of this can be that it causes bias in our data. As consequence we can have bias in our model as well. However, the number of missing data points in the several variables is really small proportionally. Therefore, we do not see it as a problem to use kNN imputation. Unless otherwise stated, we determine the $k = 5$ nearest neighbors for imputation. Note that we do not use kNN imputation for missing WOZ values, even tough we mention it as one of the variables with missing data points in the beginning of this subsection. It is our response variable and therefore we want to predict it from our model.

4.2. Regression diagnostics

After involving all explanatory variables, we check the regression diagnostics mentioned in the previous chapter. This will give us insight into the accuracy of the model. It helps us in deciding which steps we have to take next. When we started we did not had access to all variables. Over time we extended the model and tried out for example different transformations. We report here only the final results.

After fitting a model we study the added-variable plots and the component-plus-residual-plots. These plots can be found in the Appendix. For our analysis we split the data into a training and test set with a ratio 75% – 25%. The training set we use to refine the model. With the test set we try to measure how accurate the predictions and outcomes are. Furthermore, we remove outlying properties. Examples are properties with a small living space and large WOZ-value. We also set a lower bound for the WOZ-value of the properties we consider to left out possible residential garage properties.

4.2.1. Variable selection and collinearity

The first thing we address is the selection of variables. The main diagnostic we use is collinearity. We already discussed the consequences in Subsection 3.5.4. Two quantitative variables are perfectly collinear if there exists a linear relationship between them, so that one of them can be predicted from the other variable. This can be extended to qualitative variables as well. If one of them can be predicted from one or more other qualitative variables, we have collinearity as well. Examples are districts and neighborhoods. If we know in which neighborhood a property is located, we know the district as well. Using both of them will cause trouble. Qualitative variables that cause collinearity in the model are indicated by high values for the GVIF. Remember that this a measure for the deleterious impact on the least-squares estimate of the regression coefficient. Therefore, we cannot use both district and neighborhood.

So by looking at the VIF and GVIF (or $GVIF^{\frac{1}{2 \cdot Df}}$), we conclude that it is not possible to use all variables listed in Table 4.2. Besides district and neighborhood as variables, the distances to hospital and highway ramp have a high GVIF and cannot be used as well. This is even the case if we let one of the two out or use another (quadratic) transformation. It turns out that we also cannot use all five leefbaarometer dimensions. The following variables are omitted:

- Distance to hospital
- Distance to highway ramp
- Neighborhood
- RLBVRZ

Even tough these variables are omitted from the regression model, they are still being used for kNN imputation. Due to the omission of neighborhood as explanatory variable, we use the leefbaarometer data on neighborhood level. Therefore, we are still able to use neighborhood data in our model. The next model gives an overview of the (generalized) variance inflation factor for all variables in our model.

Table 4.3: Generalized variance inflation factor for the variables in our model for Delft.

Variable	**GVIF**	**Df**	$GVIF^{\frac{1}{2\cdot Df}}$
Construction year (cy)	66.51	7	1.35
Distance to supermarket	3.78	7	1.10
Distance to school	1.86	1	1.37
Distance to bus station	1.76	4	1.07
Distance to shopping mall	42.98	4	1.60
Distance to train station	22.81	4	1.48
Distance to residential boulevard	91.40	5	1.57
Property type	2.11	4	1.10
Energylabel	19.17	6	1.28
RLBFYS16	2.95	1	1.72
RLBVEI16	4.42	1	2.10
RLBWON16	5.29	1	2.30
RLBBEV16	3.27	1	1.81
Interaction between living space and district	2302.20	11	1.42

The leefbaarometer dimensions RLBVEI16 and RLBWON16 have the highest values for $GVIF^{\frac{1}{2\cdot Df}}$. Since they have only one regressor in the model, the GVIF devotes to the VIF. Looking at these values they are not far from the value 5 which is common for saying there is collinearity. However, we do not consider it as harmful for our model. Furthermore, we would like to keep these variables in the model since they provide information that the other variables do not.

We believe that there is an interaction between the living space and district. As we can see, we have an interaction term without the parental variables (living space and district). This violates the principle of marginality. However, including these terms will cause collinearity. Furthermore, we do not consider it necessary to include the parental variables.

4.2.2. Checking normality

To check the normality assumption of the residuals, we look at a QQ-plot to check whether the normality assumption of the residuals is satisfied. In particular, we study a QQ-plot of the studentized residuals against quantiles of the normal distribution.

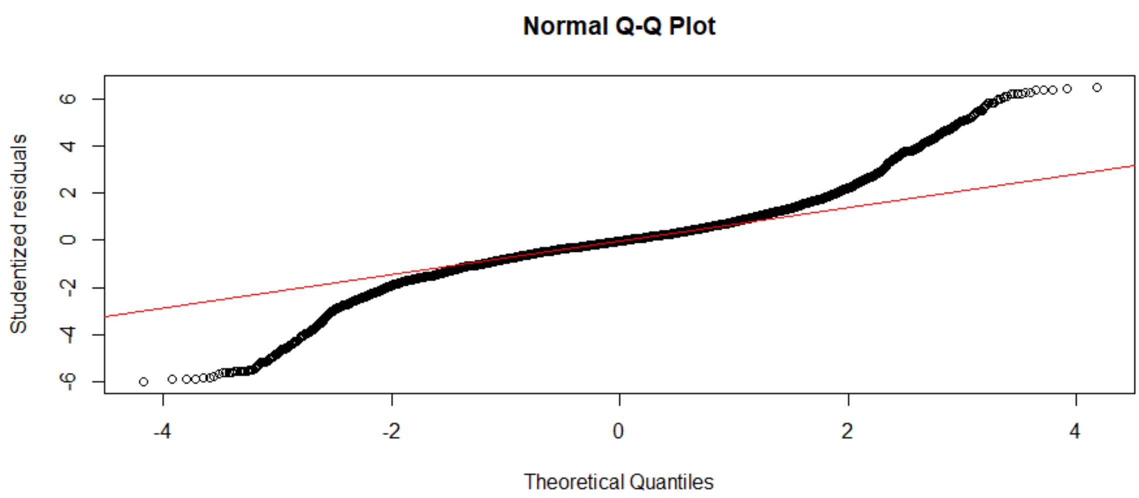


Figure 4.4: QQ-plot of the studentized residuals of the training set.

Clearly, our data is heavy tailed. Therefore the least-squares approach is not our first choice and robust estimators can be an option. We will discuss the outcome with robust estimators later.

4.2.3. Error variance
To check whether we have heteroscedasticity we plot the studentized residuals against the fitted values.

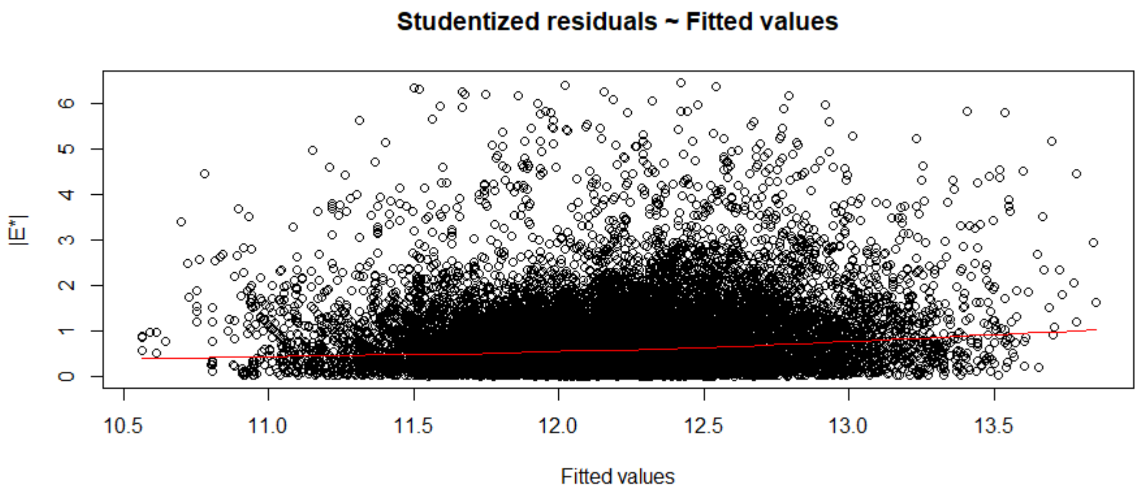


Figure 4.5: Plot of the absolute values of the studentized residuals versus the fitted values.

The red line is the loess regression line. Loess regression is a non-parametric regression method in which the predictor is constructed based on information that is derived from the data. It is often used to examine the relationship between two quantitative variables in a scatterplot. Therefore, these methods are also called scatterplot smoothers since it facilitates the interpretation of them.

As we can see in Figure 4.5, the loess regression line is slightly increasing as the fitted values are increasing. However, we do not consider it as critical to say we have heteroscedasticity, because proportionally only a small number of points are concentrated at the right side in the plot.

4.3. Model validation
We saw in the previous paragraph that we have heavy-tailed residuals. As discussed earlier, that means we have to look at robust estimators. In this paragraph we validate our model with the test data. We apply the model with the least squares estimator to the test data. We do the same with models with robust estimators. After that we compare the results.

There are various options to measure the accuracy of the model. In particular, we use two methods which we elucidate now.

The Mean Absolute Percentage Error (MAPE) is defined as

$$MAPE = 100\% \cdot \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right|. \tag{4.5}$$

It is a measure of the relative error of the predictions of the actual values.
Furthermore, we compute the ratio of the number of times a prediction is within a 10% range from the actual value relative to the total number of predictions made from the test data.

The outcomes of the two methods with the four estimators we have used can be found in Table 4.4. For the Bisquare and Huber estimator we have used the tuning constants $k = 4.685S$ and $k = 1.345S$, respectively.

Table 4.4: Validation of test data for Delft

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1084	0.6177
LAD	0.1061	0.6193
Bisquare	0.1082	0.6289
Huber	0.1077	0.6267

The MAPE we show as decimals instead of percentages. As we can see, the differences between all four estimators are really small for both validation methods. The robust estimators perform better than the least-squares estimator. Looking at the MAPE, on average there is approximately a 10.8% error of the forecasts. This is for us an acceptable outcome. However, the number of predictions that fall within a 10% range from the actual values are not that satisfying. For the regression model that is constructed with the least-squares estimator we have $\tilde{R}^2 = 0.91$.

A similar value for $\tilde{R}^2$ we also had for models with one or more of the omitted variables (see Subsection 4.2.1) included, although we did not mention it. Models can have a high $\tilde{R}^2$ even though they perform poorly regarding the predictions. Therefore, $\tilde{R}^2$ is for us not a reliable measure. We feel more familiar with diagnostics like (graphical) examinations of the residuals, variable transformations and collinearity checks. For that reason we will not give the value of $\tilde{R}^2$ anymore for the models in the subsequent chapters.

To get an impression of the model, the coefficients for all regressors are stated in a table in the first chapter of the Appendix.

5

Modeling on provincial level

After setting up a model for Delft, we want to scale up the region. Using all knowledge and insights we gained when setting up the model for Delft, we now want to apply the model on provincial level. It turns out that we cannot use exactly the same model. First we need to overcome some problems regarding the data. When this is fixed, we can apply the regression model. We start with the province Zuid-Holland, which is the province with the most properties in the Netherlands. Next we look at other provinces. We do not want to have a model for all provinces separately, since this would mean we have to run 12 different models which takes a lot of computational time. Furthermore, it will be more difficult to compare the outcomes since there are big differences in the number of properties per province. Therefore, we will sometimes combine provinces in a model. In particular, only adjacent provinces are combined to keep the diversity in a model as low as possible.

The next table gives an overview of the grouping of the provinces and the percentage of available WOZ values.

Table 5.1: Grouping of the provinces and percentage of available WOZ values

Group	**Percentage of available WOZ values**
Zuid-Holland	91.82%
Groningen, Friesland, Drenthe	78.46%
Limburg	90.86%
Noord-Brabant, Zeeland	78.83%
Gelderland, Overijssel	90.42%
Noord-Holland	88.43%
Utrecht, Flevoland	86.44%

Exact numbers and the number of properties per group can be found in Table A.14.

5.1. Zuid-Holland

After merging all different data sets as described in Chapter 2, the next step is imputation of missing data. In the model for Delft we used the k-nearest neighbor algorithm. A disadvantage of this method is that the computational time is really high and that it affords a lot of internal memory for big data sets. For every property the algorithm needs to compute the k nearest neighbors out of all properties in the data set. For Delft we had around 50000 properties. Scaling up to provincial level, we have now 1.7 million properties. It turns out that kNN imputation is not possible on provincial level and that we have to look for other imputation methods.

5.1.1. Imputation

Imputation is needed for missing energylabels, property types and leefbaarometer data. First we come up with methods to impute property types, then we look at leefbaarometer data and finally we look at the missing energylabels.

Property types

Initially, 2.6% of the types are missing. We start with finding the highest number of property types per 'pc5'. With pc5 we mean all four numbers of a postal code and the first letter, or better said the postal code without the last letter. The reason is that a pc5 area usually consists of a block of houses with the same type. The property type we have the most per pc5 is then used for imputation of the properties within this pc5. After doing this, we still have missing property types. The reason is that we do not have property types for some pc5 areas at all. The imputation continues by assigning the remaining properties the type apartment if it has a number addition as in the last column in Figure 2.1. Now only 0.02% of the labels is missing. Imputation of these labels is done based on the size of the living space and can be found in the table below.

Table 5.2: Imputation for last missing property types.

Size	**Assigned type**
living_space ≤ 90	Apartment
90 < living_space ≤ 170	Terraced house
170 < living_space ≤ 290	Semi-detached house
living_space > 290	Detached house

As we can see, the category 'House located on a corner or at the end of a terraced house construction' is excluded from the table. We do believe that it is too severely to assign this category to a property based on the data we have. Properties of this category have usually the same characteristics as terraced houses, whereas in general detached houses have a bigger living space than semi-detached houses, semi-detached houses have a higher living space than terraced houses and terraced houses have a higher living space than apartments.

Leefbaarometer

Leefbaarometer data is sometimes missing for neighborhoods. At first sight, 7.7% of the values are missing. This is a quite high number. After noticing that sometimes for a complete municipality the data was missing, we discovered that some municipalities in our data set were established since the beginning of 2019 out of other municipalities. Since the leefbaarometer data dates from 2016 and we matched the data on neighborhoods per municipalities, this data was not matched. The following changed per January the 1st of 2019.

- Municipality Hoeksche waard is a fusion of the following five former municipalities: Binnenmaas, Oud-Beijerland, Cromstrijen, Korendijk and Strijen.
- The municipality Molenlanden is a fusion of Molenwaard en Giessenlanden.
- Municipality Noordwijkerhout belongs to Noorwijk.

After correcting this, the percentage of missing data has decreased to 4.1%. The remaining missing values are filled up with the mean of the leefbaarometer values in a municipality over all properties. A disadvantage is that neighborhoods with many houses will contribute more to the mean. However, other methods will not make big differences and we think that this is an appropriate solution. We use many variables in our model and an inaccuracy in only the leefbaarometer variables will not be harmfull for the predictions.

Energylabels

For the missing energylabels (2.8% of all) we use proportional odds logistic regression. This is a type of regression where we can use qualitative variables as response. We use this regression model with the energylabels as response variable. The construction year and property types are the explanatory variables. We will not discuss this type of regression as extensively as we did with the other types in Chapter 3, because this is just for imputation purposes. More details about this type of regression can be found in [1].
Our model has a precision of 62%; this is the percentage of the labels that are predicted correctly. This is based on a test data set of size 100 000. For the train data we used 300 000 properties.

The frequency of the different energylabels and property types can be found in histograms in the Appendix.

5.1.2. Validation

We use the same validation methods as in the model for Delft. That means we look at the MAPE and the ratio of predictions that are within a 10% range from the actual value relative to the total number of predictions made from the test data. Due to the large size of the data set, running the simulations is very time consuming. It is even possible that a simulation cannot be executed due to memory issues. Therefore, we take only half of the number of properties in the province for further analysis. Again we divide this data set into a ratio 75% – 25% for the train and test data set.

We have performed different simulations with the least-squares and robust estimators for smaller samples or with properties with missing data left out. As with the models for Delft, the differences between them are small as well. The robust estimators perform slightly better than the least-squares estimator, where the bisquare performs overall the best. For this province, we thus decide to use the regression model with the bisquare estimator and give the results, based on data including imputed values, together with the least-squares estimator.

Table 5.3: Validation of test data for the model for Zuid-Holland.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1325	0.5146
Bisquare	0.1304	0.5250

Note that we do not show the QQ-plot and the plot of the absolute value of the studentized residuals versus the fitted values here. The reason is that they are almost identical to the plots for the model for Delft, see Section 4.2.2 and Section 4.2.3. What we do show here is a density plot of the estimated WOZ values. We can see that most of the properties in this province have a WOZ value between 100 000 and 400 000. The y-axis gives the values of the probability density function.

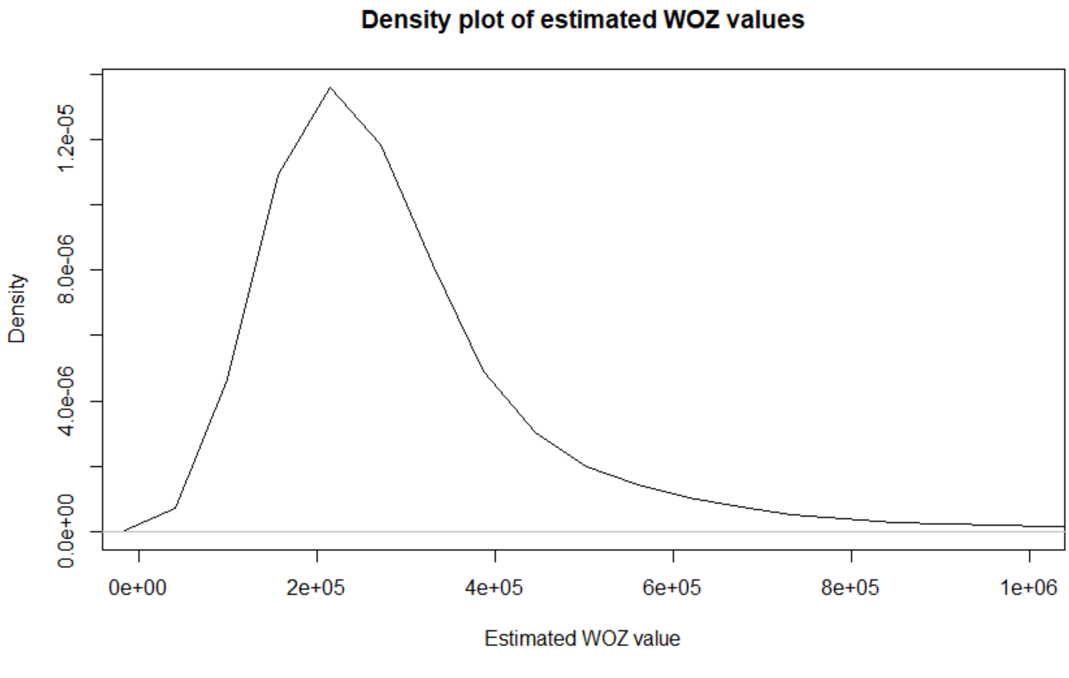


Figure 5.1: Density plot of the estimated WOZ values for Zuid-Holland.

Remarks

In the model for Delft, we used districts to keep track of the location. When modeling on provincial level, we have the problem that many cities have districts with the same name. For example, a lot of cities have a district (or neighborhood) called 'Centrum', which means centre in Dutch. To be able to still keep track of the location, we create a new qualitative variable that consists of the name of the municipality and the name of the district, separated by a underscore sign. By using factors, we can easily incorporate this variable into the model. We use this method for all other provinces in the models in this chapter as well.

Furthermore, sometimes WOZ values for all properties in a district or even municipality are missing. That means that we cannot train the model on properties in these district and municipalities. To overcome this problem, we group these district or municipalities into a nearby district. This requires accurate research on the map for each individual case. We are missing WOZ values for the municipalities Leiderdorp and Zuidplas and for the districts 'Botlek-Europoort-Maasvlakte' and 'Wijk 00 Krimpen aan den IJssel' in Rotterdam and Krimpen aan den IJssel, respectively.

5.2. Groningen, Friesland and Drenthe

In our next model we combine three provinces: Groningen, Friesland and Drenthe. Together they have almost 800 000 properties for residential use. The data imputation process is here identical with the previous model for Zuid-Holland. Also we have here missing WOZ values for seven municipalities and two districts. Finding nearby areas with appropriate sample size is even more challenging, because most of these municipalities border each other. Examples are Noardeast-Fryslân and Dantumadiel. Furthermore, we do not only have newly established municipalities by the beginning of 2019, but also since the beginning of 2018. And to make it even more complicated (regarding data cleaning), cities of a municipality are added to different municipalities in these changes. We list these changes below.

Changes by January the 1st of 2019:

- Municipality Noardeast-Fryslân is a fusion of the following four former municipalities: Dongeradeel, Ferwerderadiel en Kollumerland and Nieuwkruisland.
- Ten Boer and Haren are added to Groningen.
- Westerkwartier is a fusion of the former four municipalities Grootegast, Leek, Marum en Zuidhorn and four cities of the municipality Winsum: Ezinge, Verspreide huizen Ezinge, Feerwerd en Garnwerd.
- The rest of the cities of Winsum, together with Bedum, De Marne, Eemsmond are grouped into the new municipality Het Hogeland.

Changes by January the 1st of 2018:

- Westerwolde is a fusion of Bellingwedde and Vlagtwedde.
- Midden-Groningen is a fusion of Hoogezand-Sappemeer, Slochteren and Menterwolde.
- Waadhoeke a fusion of the former three municipalities Franekeradeel, het Bildt, Menaldumadeel (Menameradiel) and four cities of the former municipality Littenseradeel: Welsrijp, Baijum, Winsum en Spannum.

The list above is related to the changes we made in order to have less missing leefbaarometer data. If more changes were made regarding to the grouping of municipalities by governmental institutions, but they are not in the list above, it means there is initially no leefbaarometer data available at all. For example, all other cities of the former municipality Littenseradeel are added to Leeuwarden or Súdwest-Fryslân, but they are not mentioned in the list above since there is no leefbaarometer data available for the neighborhoods in these cities. More details about the regrouping of municipalities can be found via [30] and [31]. Remember that leefbaarometer data imputation is done after adapting the changes listed above.

Now we look at the outcomes of the regression model with the different estimators. We get the following results.

Table 5.4: Validation of test data for the model with Groningen, Friesland and Drenthe.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1447	0.4806
LAD	0.1447	0.4808
Bisquare	0.1444	0.4871
Huber	0.1443	0.4858

We choose here for the model with the bisquare estimator. Notably is that the results are less good than the model for Zuid-Holland. These three provinces vary from most other provinces due to their countryside structure, the relatively large number of villages and the low number of properties. There are relatively more farmhouses with large lot surfaces compared to Zuid-Holland.
Large variety in property types and property structures will lead to a less accurate model and thus a higher MAPE as we see from the table.

5.3. Other provinces

For all other provinces we do the same as in the previous two models. We will not show QQ-plots or other plots, since they are every time almost the same. Here we will only give the table with the validation of the test data, supplemented with some comments. The estimator we use is in bold letters.

Table 5.5: Validation of test data for Limburg.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1428	0.4880
LAD	0.1428	0.4879
Bisquare	0.1425	0.4924
Huber	0.1424	0.4911

Table 5.6: Validation of test data for Gelderland and Overijssel.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1397	0.5188
LAD	0.1397	0.5189
Bisquare	0.1403	0.5232
Huber	0.1397	0.5222

Table 5.7: Validation of test data for Noord-Brabant and Zeeland.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1299	0.5384
LAD	0.1299	0.5390
Bisquare	0.1291	0.5455
Huber	0.1292	0.5436

Table 5.8: Validation of test data for Utrecht and Flevoland.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1270	0.5407
LAD	0.1270	0.5407
Bisquare	0.1268	0.5476
Huber	0.1266	0.5464

Table 5.9: Validation of test data for Noord-Holland.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1287	0.5572
LAD	0.1298	0.5579
Bisquare	0.1300	0.5697
Huber	0.1292	0.5677

As was the case for Zuid-Holland, we encountered simulation problems due to memory issues for some other provinces as well. Therefore, train and test data sets were constructed after taking only half of the number of properties in the original data set. We did this for three models: Gelderland and Overijssel, Noord-Brabant and Zeeland, and Noord-Holland. While doing this, we made sure that every combination of municipality and district was represented in the train data set such that the model was not unknown with a specific combination of municipality and district and predictions could be made for all properties in the original data set.

Missing data

Here we encountered another problem regarding missing data. For some municipalities leefbaarometer data is unavailable. This holds for the following twelve municipalities: Amtelveen, Blaricum and and Laren in Noord-Holland, de Bilt and Eemnes in Utrecht and/or Flevoland, Bergeijk, Bernhaze, Cuijk, Meierijstad, Mill en Sint Hubert and Tilburg in Noord-Brabant and/or Zeeland and Wageningen in the model for Gelderland and Overijssel.

Remember that we filled up missing leefbaarometer data in neighborhoods with the mean of the leefbaarometer in the municipality. Missing data for a complete municipality means we cannot compute the mean at all. We overcome this problem by using the mean of the leefbaarometer of another nearby or similar municipality as imputation for the missing leefbaarometer data in a whole municipality. For example, since Laren and Blaricum have on average high WOZ values, leefbaarometer data of Bloemendaal is used as imputation for these two municipalities, since Bloemendaal has on average a high WOZ value as well in the province Noord-Holland [12].

5.4. Summary

We give an overview of the best results per model.

Table 5.10: Best results for each WOZ model

Model	Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Zuid-Holland	Bisquare	0.1304	0.5250
Groningen, Friesland, Drenthe	Bisquare	0.1444	0.4808
Limburg	Bisquare	0.1425	0.4924
Noord-Brabant, Zeeland	Bisquare	0.1291	0.5455
Gelderland, Overijssel	Huber	0.1397	0.5222
Noord-Holland	Huber	0.1292	0.5677
Utrecht, Flevoland	Bisquare	0.1268	0.5476

Compared to Delft, the results are less accurate. This is normal, since we considered one or more provinces instead of only one municipality.

We see that the outcomes for the different models are close to each other. We already expected that for Groningen, Friesland and Drenthe the inaccuracy would be the highest due to the diversity in the landscape structure. The MAPE for Zuid-Holland and Limburg are remarkably high, especially since these are the two groups for which the highest percentages WOZ values are available.

What we do know is that the WOZ values determined by the municipalities are not always correct or realistic. With this we do not mean a small deviation, but serious mistakes. We have seen that sometimes in our data set that there are many properties in a postal code area with a WOZ value of around 1 000 000. However, the predicted values from our model are around or even below 100 000. Looking at the other variables we see that these properties are apartments and have living spaces of size 30-40 m^2. This sounds like, proportionally, really expensive apartments. A further search on the internet shows that all these properties are located in one large flat/apartment building, where it seems impossible that all these apartments are that expensive. What we also noticed is that most of these apartments have the same house number, but a different number addition. Furthermore, we also have a property with the same house number as the apartments, but without number addition. This property also has a living space above 1000 m^2. The WOZ value is the same as all the other apartments. When we search for this property, we do not get a specific apartment in the building.

What here probably happened is that this last property we described is considered as the whole flat building. Then the corresponding WOZ value is reasonable. However, this WOZ value is also used for all the apartments inside the flat building which is almost surely a mistake.

This is one example we encountered in our data set, but we have seen more of such apartment buildings in our data set. We also noticed that there are retirements homes and communities that consist of many small apartments which have WOZ values of over a million euros. Remember that we used the BAG data set to only select properties for residential use. However, apparently properties in a retirement home or community are regarding the BAG data set classified as a residential property, which could be disputable. We tried to leave as many as we can out in the train and test sets, but sometimes some of them slip through the selection due to the large variety.

We have seen this problem with incorrect WOZ values in especially the two biggest municipalities in Zuid-Holland. But also in other provinces we have seen examples of this problem. Such a large difference between the predicted value and the original WOZ value in the data set surely will have a substantial contribution to the MAPE.

Looking at the results in Table 5.10 for the MAPE, apart from Groningen, Friesland and Drenthe and Limburg, the results are acceptable. The results in the last column are not satisfying, but not much less accurate than for Delft. Apparently, this is what we will get by using these regression models. Remember that these models are a tool to compute the final predictions since the predictions with the models used in this chapter will be used in the next chapter about asking prices. Much more can be said in the next chapter.

As we mentioned in the last sentence of Section 1.2, the models we have set up in this chapter also helps us in finding incorrect or outlying values estimated by the municipalities. This prevents us by using wrong estimates for the asking price models in the next chapter. The illustration we have given in this section about the incorrect WOZ values for apartments in one building is a clear example.

6
Asking price models

In this chapter we set up asking price models. We try to find the link between asking prices and the predicted WOZ values. This should provide us with the ability to determine the market value of properties. We use the same grouping for the provinces as in the previous chapter.

Variable selection

For an asking price model we want to use multiple regression as well. We use the sale prices as response variable. Next we have to decide which explanatory variables we want to use. Of course, one of them is the predicted WOZ value which we denote by $W\hat{O}Z$. Next we want a variable that takes into account the location. Furthermore, the data set with asking prices also provides us with a date of when a property has been put on sale. This we would like to incorporate as well. Hereafter, we can choose some of the previous used variables. We have access to asking prices of around 430 000 properties in the Netherlands. For all these properties we have almost half a million asking prices. The number of asking prices per group of provinces can be found in Table A.14. It is very unlikely that we have data for a property in every city or municipality, so we cannot set up models with cities or municipalities as categories. Therefore, neighborhoods, districts, cities and municipalities are ruled out as explanatory variables. What we do have are postal codes. By considering only the numbers, these reach out over a large region and can be used as categorical explanatory variable. So the postal code with only the numbers, which we call 'pc4' (in Subsection 5.1.1 we already introduced 'pc5'), will be our second explanatory variable. It turns out that sometimes we cannot use pc4 because of too many missings and need to reduce it to pc3 or even pc2 which are only the first three or two numbers of the postal code, respectively. This will impair the accuracy of the predictions, since we can less precisely keep track of the location.

The date of when a property has been put on sale, we would like to incorporate as numerical variable, such that the differences between consecutive days is almost equal. We have the day, the month and the year. Let X_3 be the explanatory variable for the date. Then we compute X_3 as

$$X_3 = year + (month - 1) \cdot \frac{1}{12} + (day - 1) \cdot \frac{1}{365}. \tag{6.1}$$

We take years, months and days as numerical values. For example, 14 June 2018 will give

$$X_3 = 2018 + (6-1) \cdot \frac{1}{12} + (14-1) + \frac{1}{365} = 2018.4523.$$

The advantage of this formula is that future asking prices can be computed as well. A disadvantage is that differences between the last day of a month and the first day of a subsequent month are not equal to the differences between two consecutive days in the same month. However, the differences are small and we do believe that this is a good approximation.

At last we have the option to incorporate some of the explanatory variables that were used in the models in the previous chapter. From these models, we know which variables contribute the most to $W\hat{O}Z$. Since there is a high correlation between asking prices and $W\hat{O}Z$, we expect the same behaviour of the explanatory variables when we will incorporate them in this asking price model. By looking at the regression coefficients, the most contributing variables were the living space, the combination of the municipality and district, the property type and sometimes the leefbaarometer. The latter depends on the value of the leefbaarometer itself which is usually of order 10^{-2} and not far from zero, but has a coefficient of order 10^{-1}. Although the model for Delft was used as a tool to start with, the added-variable plots for influential data in the Appendix can also be checked for the influence of the distinct variables since all the models on provincial level in the previous chapter are based on this model and are very similar. Based on these plots, the influence of the leefbaarometer function is almost zero like most of the variables. However, in the models on provincial level the coefficients are a bit different. For instance, for Delft the coefficient for the leefbaarometer variable RLBBEV16 equals $2.22 \cdot 10^{-1}$, whereas for the province Zuid-Holland it is equal to $7.00 \cdot 10^{-2}$. This can make a difference in the model.

The living space cannot be used, since it will lead to collinearity due to the high correlation with $W\hat{O}Z$. As discussed above, municipalities and districts cannot be used as well. We have simulated models with the property types, but unfortunately the hierarchical structure between the property types completely disappeared. For example, the regression coefficient for semi-detached houses was lower than the coefficient for terraced houses. We do believe that it is better to leave property types out. Ultimately, we have the leefbaarometer. This says something about the location. For this we already have the pc4 categories. Furthermore, leefbaarometer data is not always available and we had to apply imputation techniques as we saw in the previous chapter. For pc4 categories no imputation is necessary since all postal codes are available. Using both variables can also lead to possible collinearity, since for a certain pc4 category the leefbaarometer values are almost all the same. Therefore, we do not consider the leefbaarometer variable as meaningful in the model for asking prices.
In the end, we do only have a reference value of a property with $W\hat{O}Z$ and want to investigate the relationship with the asking price.

Having clarity about the explanatory variables, we can now define our regression model. We use our knowledge of variable transformation from the previous chapters. The asking price model is defined as

$$log(Y) = \alpha + \beta_1 \cdot log(X_1) + \beta_2 \cdot X_2 + \beta_3 \cdot X_3, \tag{6.2}$$

where

- Y is the asking price in euros,
- X_1 is the estimated WOZ value in euros,
- X_2 is a categorical variable for pc4 (sometimes pc3 or pc2),
- X_3 defines the date of when a property was put on sale as quantitative variable.

Clearly, inaccurate estimates for the WOZ value will lead to poor estimates for the asking prices. Sometimes we need to take the actual WOZ instead of the estimated one. If the estimated WOZ value deviates more than 20 % from the actual WOZ, we decide to replace the estimate by the actual value. For example, for Zuid-Holland we have to do this for 17% of the estimated WOZ values. We note that when we replace an estimate by the actual value, we want to exclude possible wrong WOZ values provided by the municipalites from this operation. The example in Section 5.4 with apartments in a flat building that have probably the WOZ value of the whole building are excluded from the operation.

We now proceed with the distinct regions.

6.1. Groningen, Friesland and Drenthe

We start with combining Groningen, Friesland and Drenthe into one model.

First we look at a component-plus-residual plot to look whether we can find any non-linear patterns.

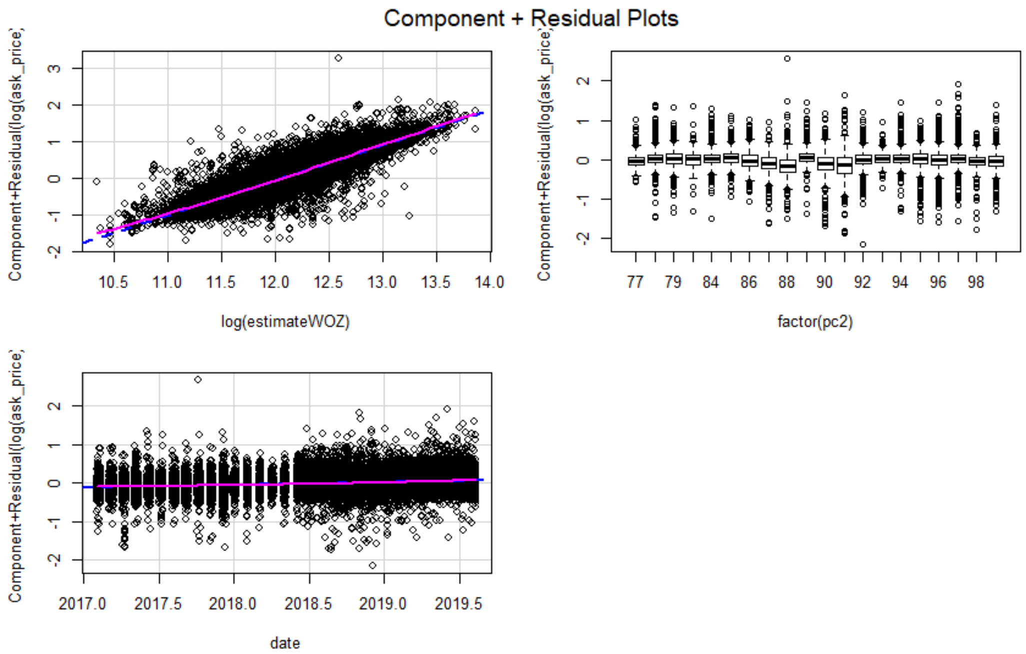


Figure 6.1: Component-plus-residual plot for asking price model for Groningen, Friesland and Drenthe.

This shows us that no further action is required regarding variable transformations or the selection of variables.
As in the models in the previous chapters, we look again at a QQ-plot and a plot of the studentized residuals versus the fitted values.

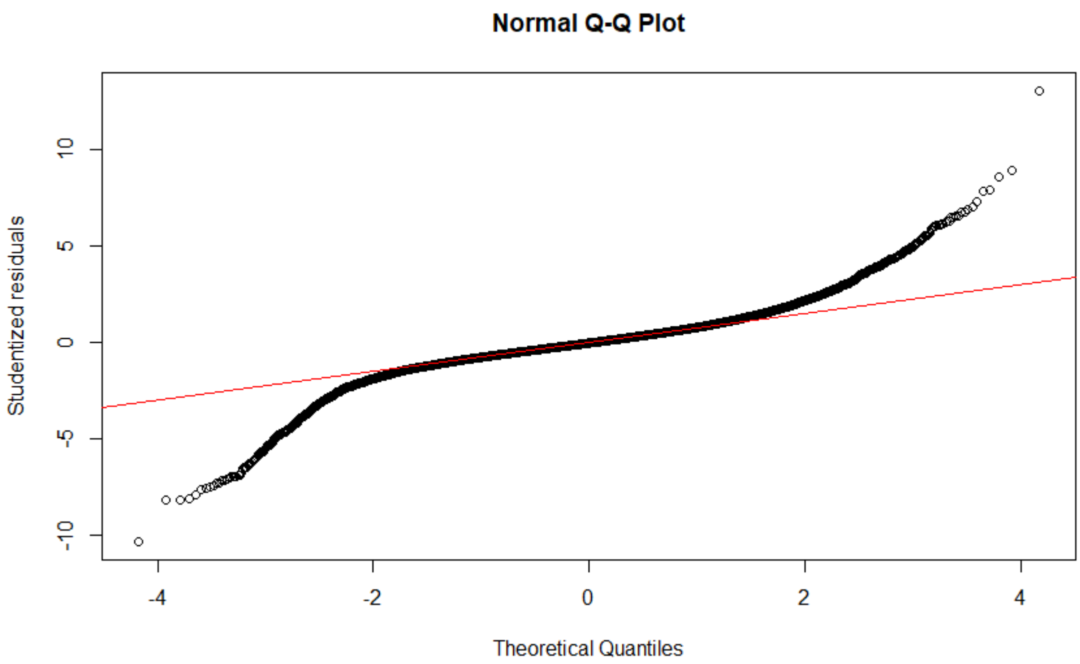


Figure 6.2: QQ-plot of the studentized residuals of the training set for asking prices.

As expected, our data is heavy tailed.

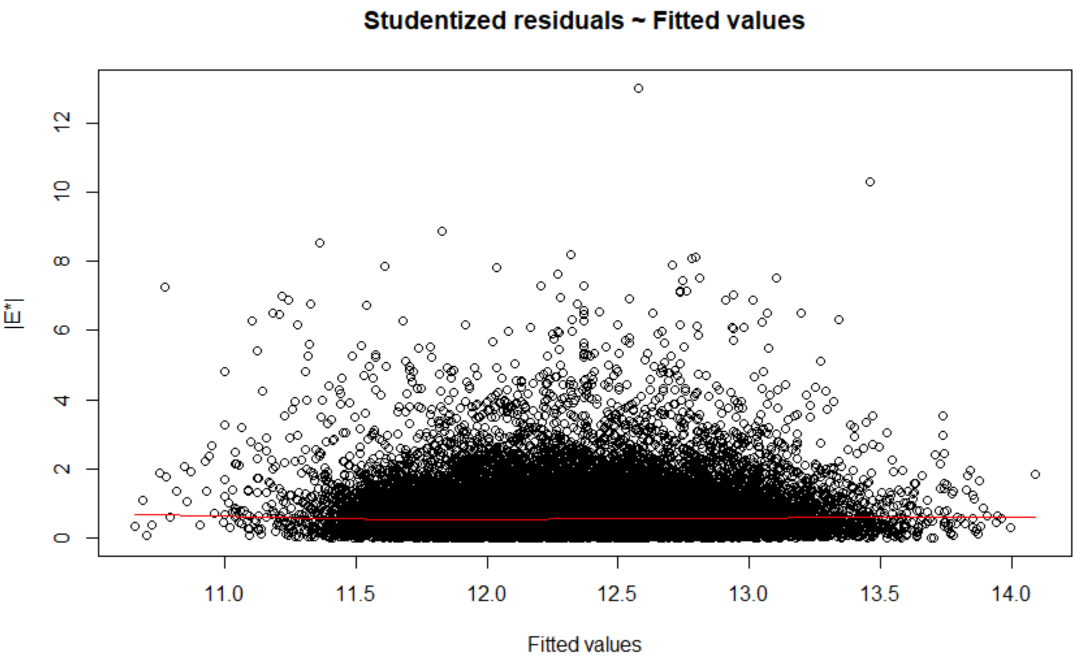


Figure 6.3: Plot of the absolute values of the studentized residuals versus the fitted values.

The loess regression line remains almost constant. Therefore we conclude that we do not have heteroscedasticity.

The following table gives the results for the well-known four estimators.

Table 6.1: Validation of test data for asking price model with Groningen, Friesland and Drenthe.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1502	0.4671
LAD	0.1502	0.4668
Bisquare	0.1493	0.4698
Huber	0.1494	0.4700

Here we prefer the Huber estimator over the other estimators. Two pc2 categories have no asking price, namely '86' and '87'. We group them into category '85' and '88', respectively.

6.2. Zuid-Holland

For this province we have almost twice as much asking prices available compared to the provinces that are used in the previous section. Fortunately, we can use here pc4 to keep track of the location. In the end, 28 pc4 categories have no asking prices. Therefore we group them into a nearby area. For example, properties in Moerkapelle and Zevenhuizen are placed into a pc4 category that belongs to Zoetermeer. Or the categories '2921' to '2926' from Krimpen aan den Ijssel and the categories '2911' to '2914' from Nieuwerkerk aan den Ijssel are grouped into a category from Capelle aan den Ijssel.

We will not show the QQ-plot and plot of the studentized residuals versus the fitted values, since they are for every model almost identical. We immediatly look at the results of the validation of test data.

Table 6.2: Validation of test data for asking price model for Zuid-Holland.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1206	0.5298
LAD	0.1190	0.5346
Bisquare	0.1184	0.5376
Huber	0.1187	0.5364

We will stick here to the bisquare estimator since this one performs the best. Now we give a plot of the asking prices from the test data set versus the predictions with the bisquare estimator.

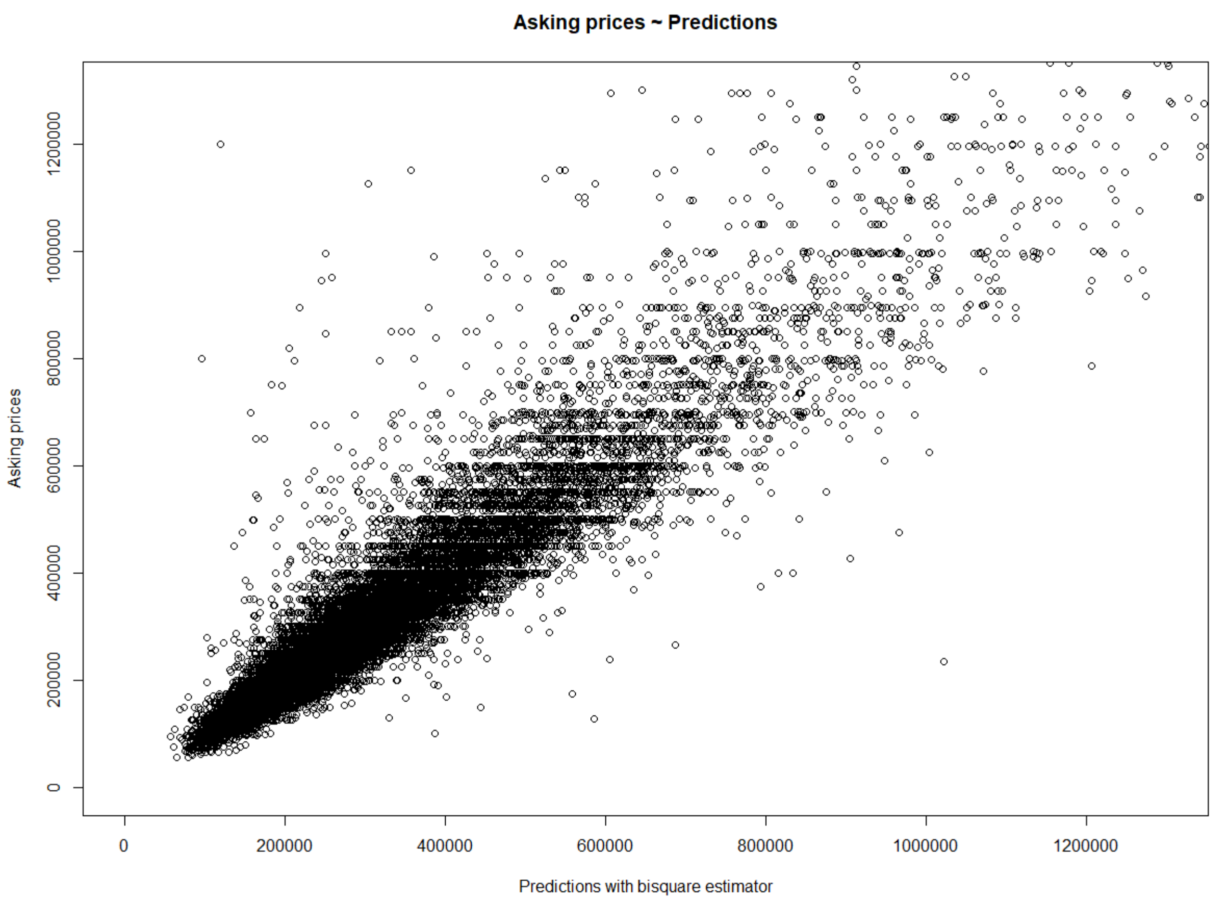


Figure 6.4: Plot of the asking prices in Zuid-Holland versus the predictions of the asking prices with the bisquare estimator.

Of course, we expected better results here than for Groningen, Friesland and Drenthe, due a regular structure (mainly cities and less villages and rural areas.), the higher number of available asking prices, and the possibility to keep better track of the location by using pc4 instead of pc2. We have here more than 500 categories for pc4, whereas we had barely 20 when using pc2 for Groningen, Friesland and Drenthe.

We finish this section with a plot of the predicted asking prices verses the estimated WOZ values. This gives us an idea about the relationship between them.

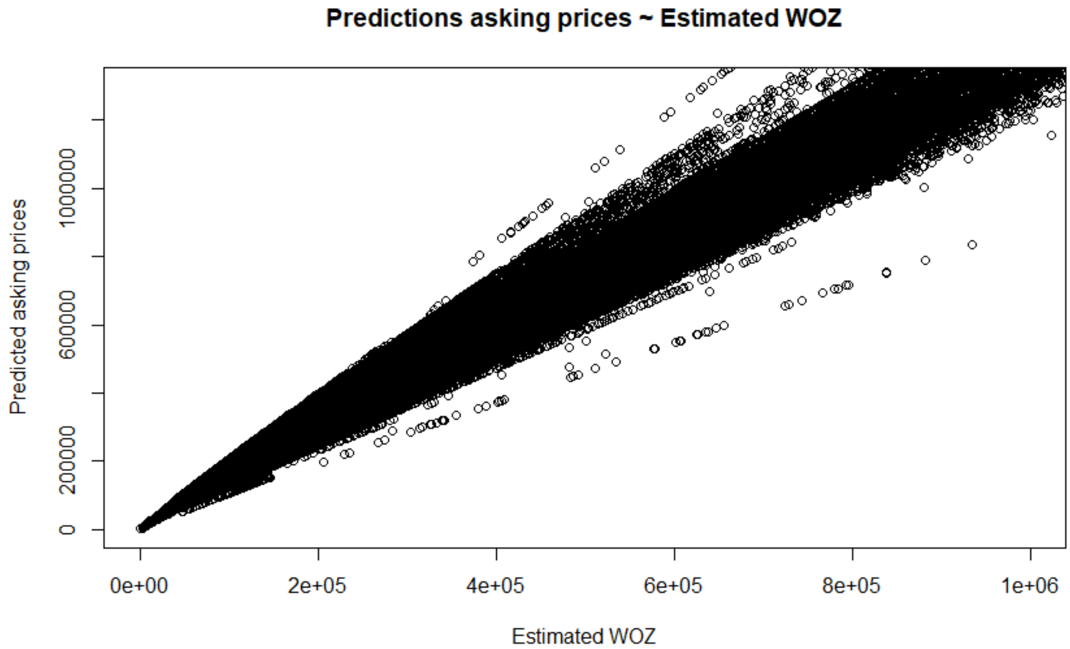


Figure 6.5: Plot of the predicted asking prices in Zuid-Holland versus the estimated WOZ with the Huber estimator.

It is common to have less accuracy for higher values of the estimated WOZ, since most of the data falls within the range of 100 000 to 400 000 as we saw in the previous chapter in Figure 5.1. As consequence, the data within this range contributes the most to the model. Furthermore, the large variation for higher estimated WOZ values can also be explained by the high number of categories for pc4.

6.3. Other provinces

For all other provinces we do the same as in the previous two asking price models. We will not show QQ-plots or other plots, since they look every time almost the same or show the same pattern in terms of accuracy. Here we will only give the table with the validation of the test data, supplemented with some comments and/or remarks. The estimator we use is in bold letters.

Table 6.3: Validation of test data for asking price model for Limburg.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1413	0.4735
LAD	0.1400	0.4795
Bisquare	0.1398	0.4798
Huber	0.1399	0.4802

Table 6.4: Validation of test data for asking price model for Gelderland and Overijssel.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1253	0.5249
LAD	0.1243	0.5287
Bisquare	0.1236	0.5323
Huber	0.1239	0.5309

For Gelderland and Overijssel we use pc3 categories instead of pc4.

Table 6.5: Validation of test data for asking price model for Noord-Brabant and Zeeland.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1217	0.5327
LAD	0.1211	0.5364
Bisquare	0.1205	0.5390
Huber	0.1207	0.5381

For Noord-Brabant and Zeeland we use pc3 categories.

Table 6.6: Validation of test data for asking price model for Utrecht and Flevoland.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1195	0.5634
LAD	0.1190	0.5703
Bisquare	0.1190	0.5723
Huber	0.1189	0.5714

Table 6.7: Validation of test data for asking price model for Noord-Holland.

Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$
Least-squares	0.1272	0.5177
LAD	0.1264	0.5227
Bisquare	0.1250	0.5278
Huber	0.1253	0.5253

For Noord-Holland we use pc3 categories.

6.4. Summary

We give an overview of the best results per model.

Table 6.8: Best results for each asking price model

Model	Estimator	MAPE	ratio of $\hat{Y} \in [0.9 \cdot Y, 1.1 \cdot Y]$	PC category
Zuid-Holland	Bisquare	0.1184	0.5376	pc4
Groningen, Friesland, Drenthe	Huber	0.1494	0.4700	pc2
Limburg	Huber	0.1399	0.4802	pc4
Noord-Brabant, Zeeland	Bisquare	0.1205	0.5390	pc3
Gelderland, Overijssel	Bisquare	0.1236	0.5323	pc3
Noord-Holland	Bisquare	0.1250	0.5278	pc3
Utrecht, Flevoland	Bisquare	0.1190	0.5723	pc4

We already discussed in the first paragraph of this chapter the outcome for the three northern provinces. Looking at the table above, the results for Limburg surprise us the most. It is the smallest model in terms of the number of properties, it is possible to use pc4 and proportionally we do not have much less asking prices compared to other models. However, Groningen, Friesland and Drenthe excluded, it is the least good performing model. We are not fully convinced about this model, but the outcome for Limburg is not dramatic.

If we look at the other five models, the two models where we used pc4 are slightly better than the three models where pc3 was used. But the differences are minimal. It would be interesting to know what the outcomes would be if it was possible to use pc4 instead of pc3.

Looking at the results, except for Limburg, all models have our approval. The MAPE outcomes are acceptable.

7

Discussion, conclusion and recommendations

7.1. Discussion

We started this project with a literature study. The studied methods were based on sale prices over a time period over many years. Since we had no access to sale prices at all and did not have useful data for a long time period, we had to come up with an alternative approach.
By using WOZ values we had reference values that could be used to start with. Even though we used them as response variable in a regression model, our goal has never been to just raise up these values for the final predictions. Firstly, we wanted to come up with our 'own' WOZ of 2017, which is based on a model after thorough research. Secondly, we set up a model that was able to predict the asking price on a certain date of a house, where our predicted WOZ was used as explanatory variable.

All data cleaning and imputation processes required a substantial amount of time. Missing WOZ values in a district or even a complete municipality were solved by grouping the properties without WOZ value in another district of the same municipality or in a district of another municipality, respectively. This grouping was done after careful looking on a map. As consequence, it is possible that the estimates for these properties are less accurate. However, much cannot be said about it. Missing WOZ values means that we have no reference values at all. Possible inaccuracy could be resolved after running the model with asking prices. However, there we had the problem with missing asking prices in a postcal code area. So regrouping was here necessary as well.
Our experience is that missing asking prices in a postal code area is not problematic. The predicted asking prices will still approach the actual values of the asking prices of houses that are currently for sale. On the contrary, inaccuracy is more often measured when the WOZ value for a complete municipality was missing. Final predictions of asking prices are in that case often underestimated. Remember that the initial goal of this project is to come up with the current values of all houses. We do not have data of these current values and therefore cannot compare our final predictions. These statements are based on a simple search of houses that are currently for sale.

When setting up the models, we took into account various regression diagnostics. Most of them were based on analysis of the (studentized) residuals. For instance, we checked for normality, linearity and heteroscedasticity by using graphical examinations. Furthermore we applied variable transformations and checked for collinearity. The latter helped us in the decision making of which explanatory variables could be used. Having checked and applied these diagnostics means more to us than outcomes like 50 – 60% for the ratio of the number of times a prediction is within a 10% range from the actual value relative to the total number of predictions made from the test data. We already mentioned that we were not interested in the adjusted R^2, since models can have a high $\tilde{R}^2$ even though they perform poorly regarding the predictions.

Regarding the validation of predicted values we used the MAPE. Of course there were more options like the mean squared error, but we think that MAPE was an appropriate validator due to its simplicity and the context of this project. It is easy to interpret, all values are above zero and it gives an average deviation from the value of the response variable in percentages.

The asking price data we received two months before finishing this thesis. The data of all provinces other than Zuid-Holland, Groningen, Friesland and Drenthe we received 5-6 weeks before finishing this thesis. Therefore, much time was not available to try different things. For instance, combining other provinces in a model than we did.

We made predictions of the asking price on a certain day. From these predictions, the people at TJIP will derive the value of a house.

7.2. Conclusion

We have used 14 different models to come to a final prediction: seven models for the prediction of the WOZ of 2017 and seven models for the prediction of the asking price. We did this by using regression analysis combined with various diagnostics. As we saw, housing data is heavy tailed and therefore all our models are based on robust estimators.

Regarding the data, we considered various information. We look at the characteristics of the house itself, information about the place where it is located and the surroundings and incorporated the trend on the housing market by using asking prices. The latter is mainly an effect of the economic situation (e.g. financial power, willingness of banks to provide mortgage loans, mortgage interests) and the relationship between supply and demand. It is a good representation of the currently over strained housing market.

In Chapter 4 we set up the model for Delft which was meant as tool for the models on provincial level. The results for Delft regarding the MAPE were acceptable. Looking at the ratio of predictions that deviate at most 10% from the actual values they are not convincing, but apparently this is what we can get with these models. The inaccuracy would decrease by scaling up to provincial level, which is a logical consequence. We saw in Chapter 5 that the outcomes for the models were acceptable as well, where we give Limburg the benefit of doubt since the results were not dramatic. Moreover, the same holds for the asking price models in Chapter 6.

Overall, we think that the models surely can be used for valuation of residential real estate. Due to the large variety it is difficult to account for all types and variations, especially in the northern provinces. To answer the second question in the first sentence of Chapter 1, which asked which factors contribute the most to the actual value of a house, we can say the following. Based on the added-variable plots and regression coefficients the main factors for the value are the living space, property type and district where the house is located. Using neighborhoods instead of district would probably be even better since this is on smaller scale, but we can only say something about districts since that is what we used. This is based on the WOZ models.
For the asking price it is important to measure the trend on the market for distinct locations, which are preferably pc4 areas.

Due to the subjectivity in the value of a house it does not seem possible to come up with an exact value. However, by coming up with a prediction of the asking price, we account for the inaccuracy and are close to the market value.

7.3. Recommendations

Here we give some recommendations that can be taken into account to improve the results when using the models in the future. Most of these recommendations strike back at data collection.

First of all we see a difference in the outcomes of the validation results in the models for different regions which is among others a consequence of the variety of the housing structure. As we mentioned earlier, in the northern provinces we have many farm houses of average size with a large lot size of hundreds or thousands hectares. Our data set does not have lot size data and therefore the predictions for these type of properties are less accurate. Also, from the data we do not even know whether there is a lot. Therefore, obtaining lot sizes would be helpful in making predictions for these type of properties.
Data of other special features of a property that can raise the price, for instance swimming pools and garage

boxes, can be valuable as well. We did not account for these features, since we had no corresponding data available.

We used the same grouping for property types as the Dutch cadastre. This was the most convenient option, since the data set with temporary energylabels provides property types with this grouping as well. As consequence some special types of properties which are very different from the five types we used, are grouped into categories where they have nothing in common with. Townhouses and canal houses are good examples. As consequence, townhouses are usually underestimated in the final predictions. It would be helpful to distinguish between these special types, such that more accurate predictions can be made. At the same time we do realize that obtaining this data is not easy.

Regarding the mathematics, we used four different regression estimators. In particular, we used two robust estimators: Tukey's bisquare and Hubers estimator. Another option is to use the Hampel estimator, which is an M-estimator as well.

Finally, it would be interesting to know what outcomes we would get if pc4 could be used for the asking price models where it was not possible to use them. This is a matter of obtaining more data over a longer time period. Also, other groupings for the provinces in the models, or using no grouping at all but simulating each province separately could be considered.

A

Appendix

A.1. Model for Delft

A.1.1. Variable transformations

Here we give an overview of the variables and the categories that were not discussed earlier.

Table A.1: Grouping of the categories for supermarket distances in Delft.

Categories for distance to nearest supermarket in Delft	
Group name	Distance (x) range in meter
0	x ≤ 100
1	$100 < x \leq 200$
2	$200 < x \leq 300$
3	$300 < x \leq 400$
4	$400 < x \leq 500$
5	$500 < x \leq 600$
6	$600 < x \leq 700$
7	$x > 700$

Table A.2: Grouping of the categories for distances to bus stations in Delft.

Categories for distance to nearest bus station in Delft	
Group name	Distance (x) range in meter
0	$x \leq 100$
1	$100 < x \leq 200$
2	$200 < x \leq 300$
3	$300 < x \leq 400$
4	$x > 400$

Table A.3: Grouping of the categories for distances to shopping malls in Delft.

Categories for distance to nearest shopping mall in Delft	
Group name	Distance (x) range in meter
0	$x \leq 500$
5	$500 < x \leq 1000$
10	$1000 < x \leq 1500$
15	$1500 < x \leq 2000$
20	$x > 2000$

Note that the grouping of distances to train stations is identical to the grouping of distance to shopping malls.

Table A.4: Grouping of the categories for distances to train stations in Delft.

Categories for distance to nearest train station in Delft	
Group name	Distance (x) range in meter
0	$x \leq 500$
5	$500 < x \leq 1000$
10	$1000 < x \leq 1500$
15	$1500 < x \leq 2000$
20	$x > 2000$

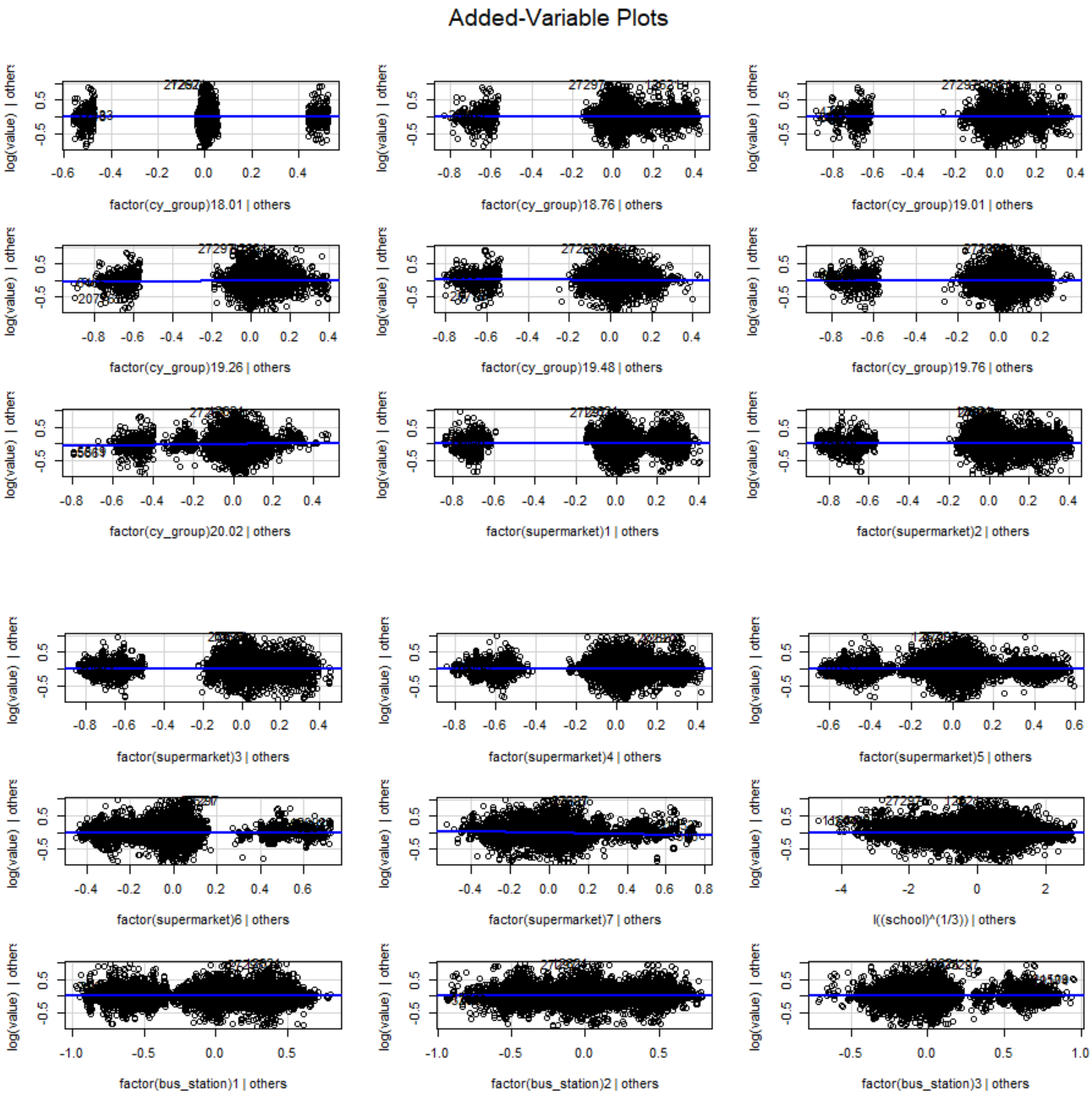
Table A.5: Grouping of the categories for distances to residential boulevards in Delft.

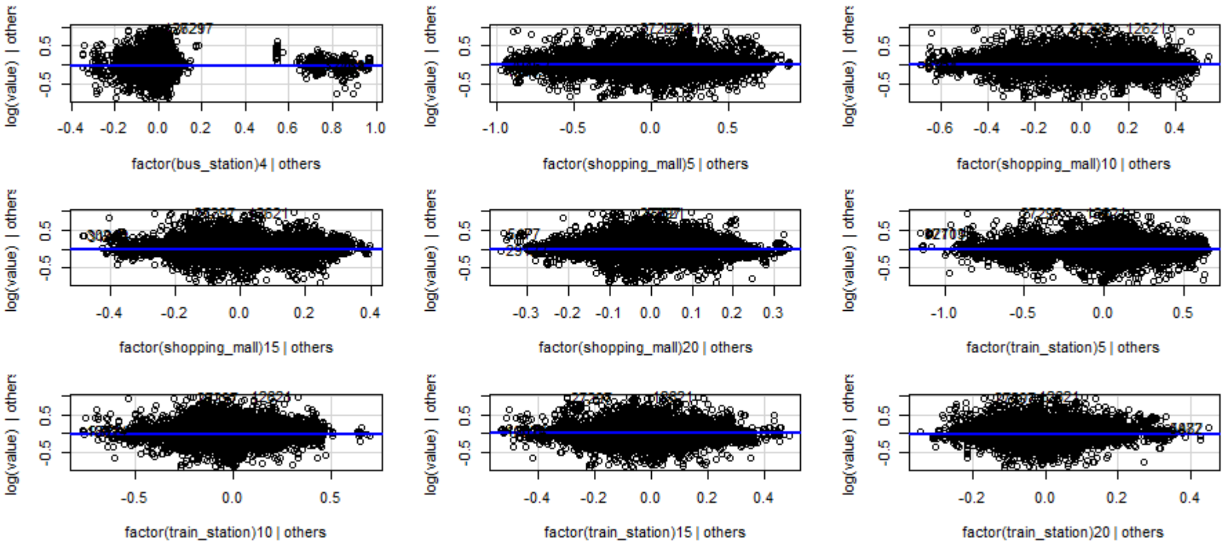
Categories for distance to residential boulevard mall in Delft	
Group name	Distance (x) range in meter
0	$x \leq 500$
5	$500 < x \leq 1000$
10	$1000 < x \leq 1500$
15	$1500 < x \leq 2000$
20	$2000 < x \leq 2500$
25	$x > 2500$

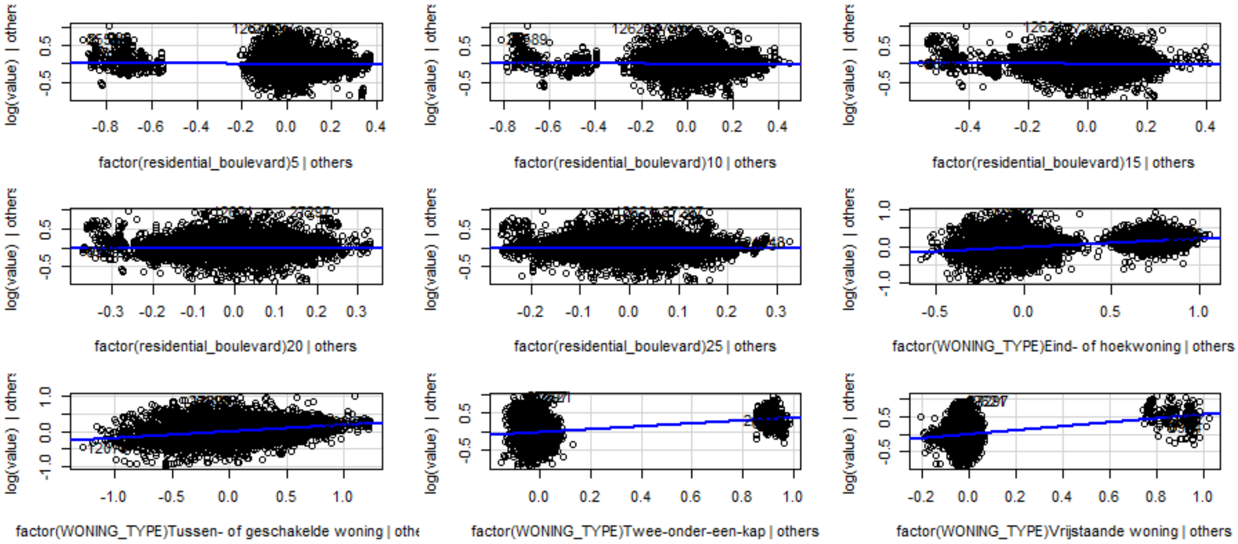
Table A.6: Grouping of the categories for energylabels.

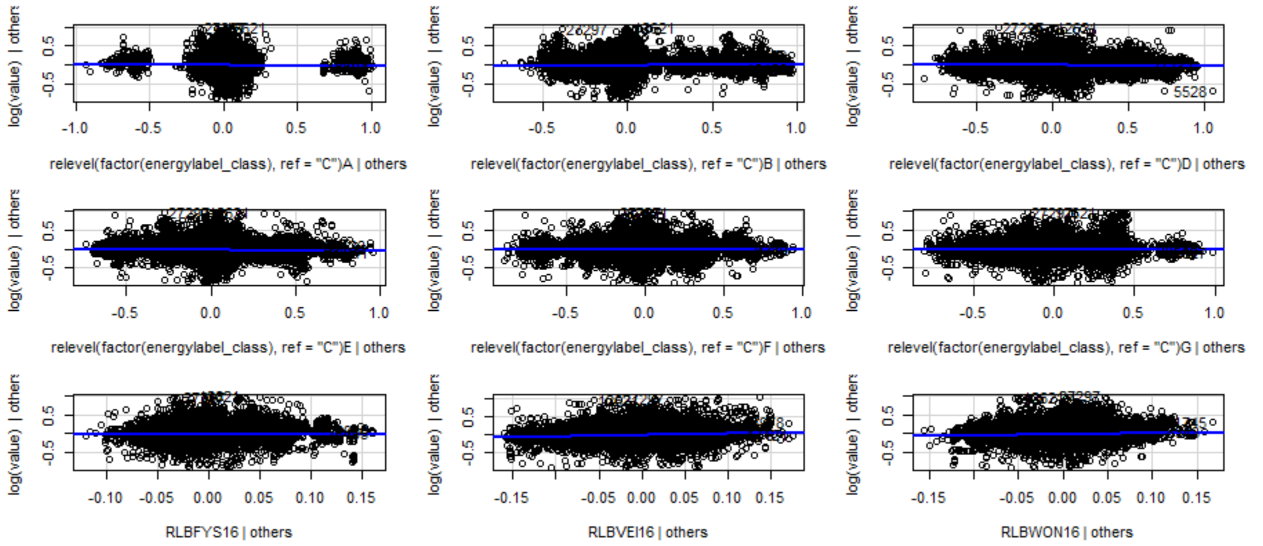
Categories for energylabels in Delft	
Group	Energylabel
A	A
A	A+
A	A ++
B	B
C	C
D	D
E	E
F	F
G	G

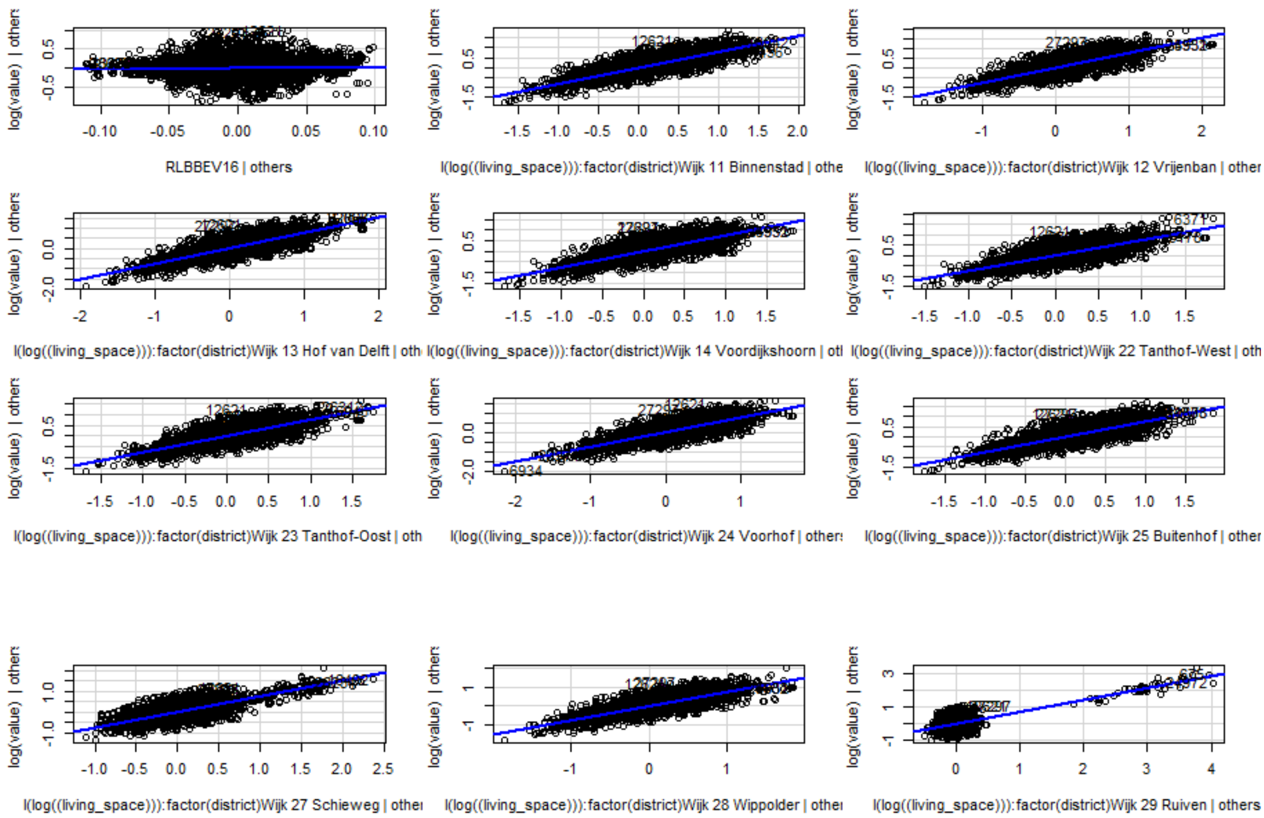
A.1.2. Added-variable plots




log(value) | others
factor(bus_station)4 | others
factor(shopping_mall)5 | others
factor(shopping_mall)10 | others
factor(shopping_mall)15 | others
factor(shopping_mall)20 | others
factor(train_station)5 | others
factor(train_station)10 | others
factor(train_station)15 | others
factor(train_station)20 | others

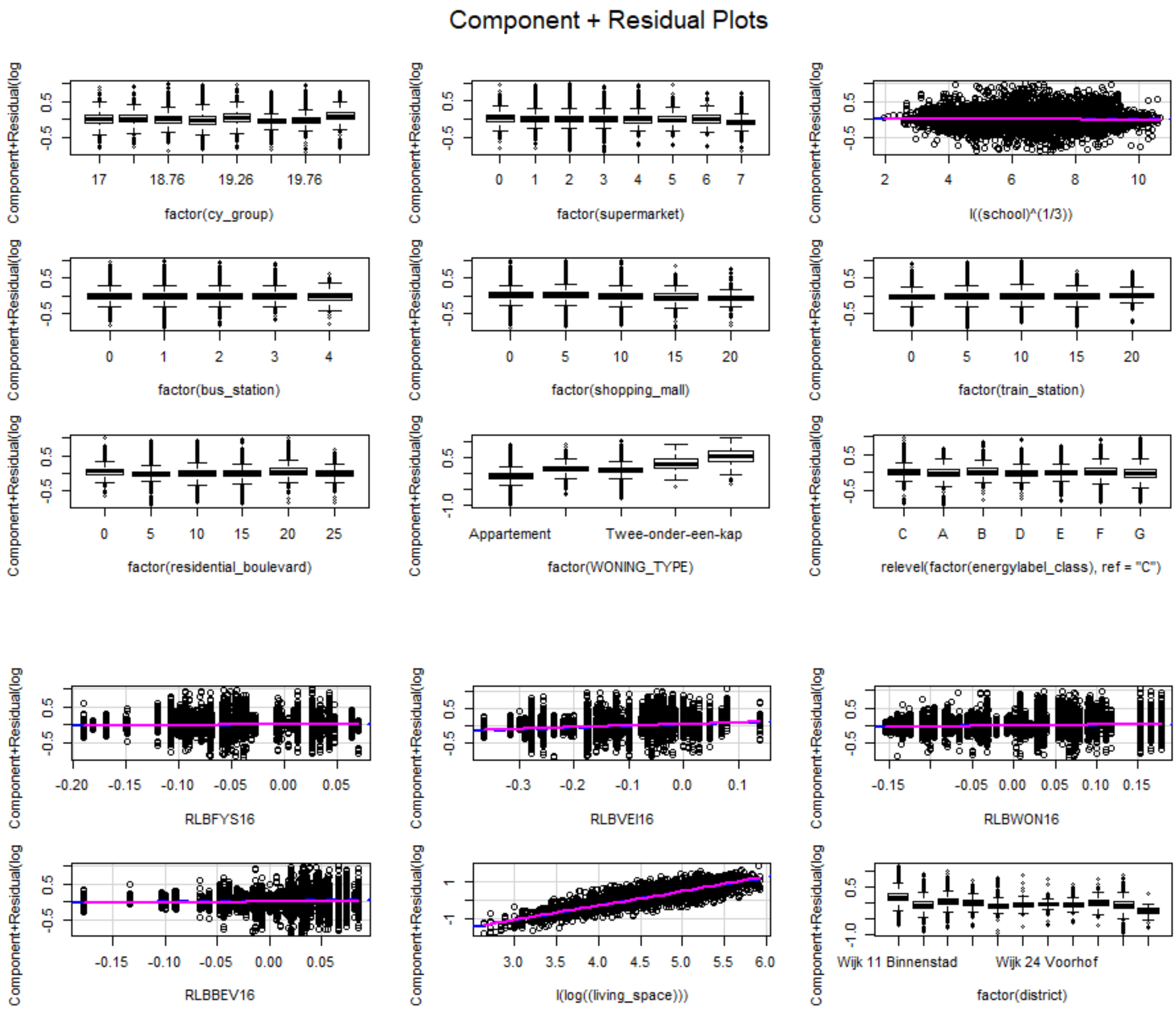

log(value) | others
factor(residential_boulevard)5 | others
factor(residential_boulevard)10 | others
factor(residential_boulevard)15 | others
factor(residential_boulevard)20 | others
factor(residential_boulevard)25 | others
factor(WONING_TYPE)Eind- of hoekwoning | others
factor(WONING_TYPE)Tussen- of geschakelde woning | othe
factor(WONING_TYPE)Twee-onder-een-kap | others
factor(WONING_TYPE)Vrijstaande woning | others


relevel(factor(energylabel_class), ref = "C")A | others
relevel(factor(energylabel_class), ref = "C")B | others
relevel(factor(energylabel_class), ref = "C")D | others
relevel(factor(energylabel_class), ref = "C")E | others
relevel(factor(energylabel_class), ref = "C")F | others
relevel(factor(energylabel_class), ref = "C")G | others
RLBFYS16 | others
RLBVEI16 | others
RLBWON16 | others
log(value) | others


RLBBEV16 | others
I(log((living_space))):factor(district)Wijk 11 Binnenstad | othe
I(log((living_space))):factor(district)Wijk 12 Vrijenban | other
I(log((living_space))):factor(district)Wijk 13 Hof van Delft | oth
I(log((living_space))):factor(district)Wijk 14 Voordijkshoorn | otl
I(log((living_space))):factor(district)Wijk 22 Tanthof-West | oth
I(log((living_space))):factor(district)Wijk 23 Tanthof-Oost | oth
I(log((living_space))):factor(district)Wijk 24 Voorhof | other:
I(log((living_space))):factor(district)Wijk 25 Buitenhof | other
I(log((living_space))):factor(district)Wijk 27 Schieweg | othe
I(log((living_space))):factor(district)Wijk 28 Wippolder | othe
I(log((living_space))):factor(district)Wijk 29 Ruiven | others
log(value) | others

A.1.3. Component-plus-residual-plots

Note: component-plus-residual-plots are not available for models with interaction terms. Therefore we give here the plots without the interaction term, but including the parental terms separately.



A.1.4. Regression model

To give the reader an idea of how the regression model looks, we give here an overview. First we give a table with the reference categories for each categorical variable. Details about the first six reference groups can be found in Table 4.1, Table A.1, Table A.2, Table A.3, Table A.4 and Table A.5.

Table A.7: Reference categories for regression model for Delft

Categorical variable	**Reference group**
construction year (cy)	17
Distance to supermarket	0
Distance to bus station	0
Distance to shopping mall	0
Distance to train station	0
Distance to residential boulevard	0
Property type	Apartment
Energylabel	C

Table A.8: Regression model for Delft

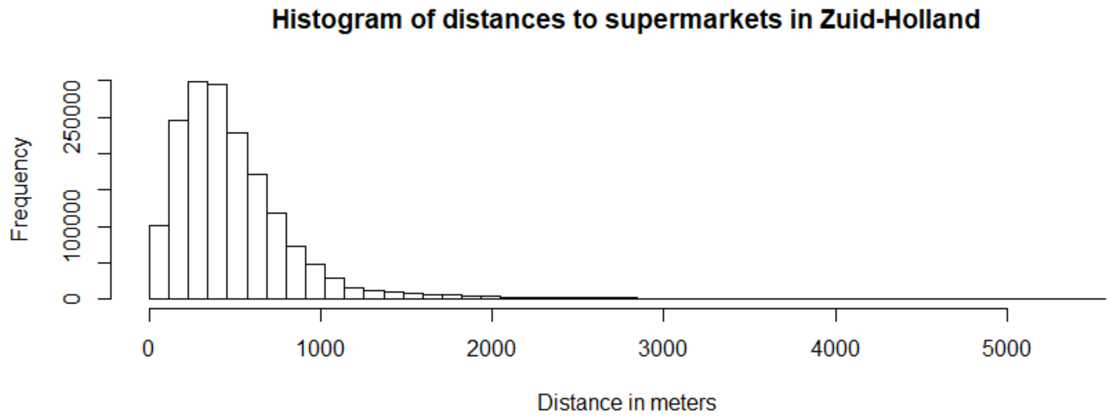
Regressor	Coefficient	Standard Error
Intercept	8.81	$1.61 \cdot 10^{-2}$
factor cy 18.01	$-3.34 \cdot 10^{-3}$	$9.33 \cdot 10^{-3}$
factor cy 18.76	$-1.09 \cdot 10^{-2}$	$8.25 \cdot 10^{-3}$
factor cy 19.01	$-2.97 \cdot 10^{-2}$	$7.96 \cdot 10^{-3}$
factor cy 19.26	$1.40 \cdot 10^{-2}$	$8.18 \cdot 10^{-3}$
factor cy 19.48	$-7.36 \cdot 10^{-2}$	$8.27 \cdot 10^{-3}$
factor cy 19.76	$-3.99 \cdot 10^{-2}$	$8.01 \cdot 10^{-3}$
factor cy 20.02	$7.45 \cdot 10^{-2}$	$9.52 \cdot 10^{-3}$
factor supermarket 1	$-8.41 \cdot 10^{-3}$	$3.69 \cdot 10^{-3}$
factor supermarket 2	$-1.10 \cdot 10^{-2}$	$3.69 \cdot 10^{-3}$
factor supermarket 3	$-1.85 \cdot 10^{-2}$	$3.80 \cdot 10^{-3}$
factor supermarket 4	$-3.42 \cdot 10^{-2}$	$3.96 \cdot 10^{-3}$
factor supermarket 5	$-1.70 \cdot 10^{-2}$	$4.54 \cdot 10^{-3}$
factor supermarket 6	$-1.21 \cdot 10^{-2}$	$5.67 \cdot 10^{-3}$
factor supermarket 7	$-1.07 \cdot 10^{-1}$	$6.13 \cdot 10^{-3}$
$(\text{school})^{\frac{1}{3}}$	$-2.51 \cdot 10^{-3}$	$8.66 \cdot 10^{-4}$
factor bus station 1	$-2.67 \cdot 10^{-3}$	$2.14 \cdot 10^{-3}$
factor bus station 2	$-9.91 \cdot 10^{-4}$	$2.46 \cdot 10^{-3}$
factor bus station 3	$-6.68 \cdot 10^{-3}$	$3.77 \cdot 10^{-3}$
factor bus station 4	$-2.05 \cdot 10^{-2}$	$7.33 \cdot 10^{-3}$
factor shopping mall 5	$8.70 \cdot 10^{-4}$	$3.58 \cdot 10^{-3}$
factor shopping mall 10	$-2.62 \cdot 10^{-2}$	$4.68 \cdot 10^{-3}$
factor shopping mall 15	$-4.73 \cdot 10^{-2}$	$6.31 \cdot 10^{-3}$
factor shopping mall 20	$-6.62 \cdot 10^{-2}$	$8.47 \cdot 10^{-3}$
factor train station 5	$8.54 \cdot 10^{-3}$	$3.54 \cdot 10^{-3}$
factor train station 10	$1.74 \cdot 10^{-2}$	$4.88 \cdot 10^{-3}$
factor train station 15	$1.42 \cdot 10^{-2}$	$6.38 \cdot 10^{-3}$
factor train station 20	$3.02 \cdot 10^{-2}$	$8.56 \cdot 10^{-3}$
factor residential boulevard 5	$-7.22 \cdot 10^{-2}$	$5.06 \cdot 10^{-3}$
factor residential boulevard 10	$-7.48 \cdot 10^{-2}$	$5.71 \cdot 10^{-3}$
factor residential boulevard 15	$-6.94 \cdot 10^{-2}$	$7.00 \cdot 10^{-3}$
factor residential boulevard 20	$-1.48 \cdot 10^{-2}$	$8.60 \cdot 10^{-3}$
factor residential boulevard 25	$-7.22 \cdot 10^{-2}$	$1.03 \cdot 10^{-2}$
factor property type terraced house	$1.92 \cdot 10^{-1}$	$2.45 \cdot 10^{-3}$
factor property type house located on a corner	$2.41 \cdot 10^{-1}$	$3.47 \cdot 10^{-3}$
factor property type semi-detached house	$3.82 \cdot 10^{-1}$	$7.56 \cdot 10^{-3}$
factor property type detached house	$5.90 \cdot 10^{-1}$	$1.02 \cdot 10^{-2}$
factor energylabel A	$-2.32 \cdot 10^{-2}$	$5.55 \cdot 10^{-3}$
factor energylabel B	$2.15 \cdot 10^{-2}$	$3.46 \cdot 10^{-3}$
factor energylabel D	$-3.10 \cdot 10^{-2}$	$2.82 \cdot 10^{-3}$
factor energylabel E	$2.73 \cdot 10^{-2}$	$3.14 \cdot 10^{-3}$
factor energylabel F	$4.77 \cdot 10^{-3}$	$3.67 \cdot 10^{-3}$
factor energylabel G	$-2.05 \cdot 10^{-2}$	$4.21 \cdot 10^{-3}$
RLBFYS16	$1.30 \cdot 10^{-1}$	$2.45 \cdot 10^{-2}$
RLBVEI16	$5.08 \cdot 10^{-1}$	$1.74 \cdot 10^{-3}$
RLBWON16	$2.91 \cdot 10^{-1}$	$2.04 \cdot 10^{-3}$
RLBBEV16	$2.22 \cdot 10^{-1}$	$2.84 \cdot 10^{-3}$
Interaction: log(living space) * factor district ‘Wijk 11 Binnenstad’	$8.03 \cdot 10^{-1}$	$2.32 \cdot 10^{-3}$
Interaction: log(living space) * factor district ‘Wijk 12 Vrijenban’	$7.52 \cdot 10^{-1}$	$2.35 \cdot 10^{-3}$
Interaction: log(living space) * factor district ‘Wijk 13 Hof van Delft’	$7.73 \cdot 10^{-1}$	$2.29 \cdot 10^{-3}$

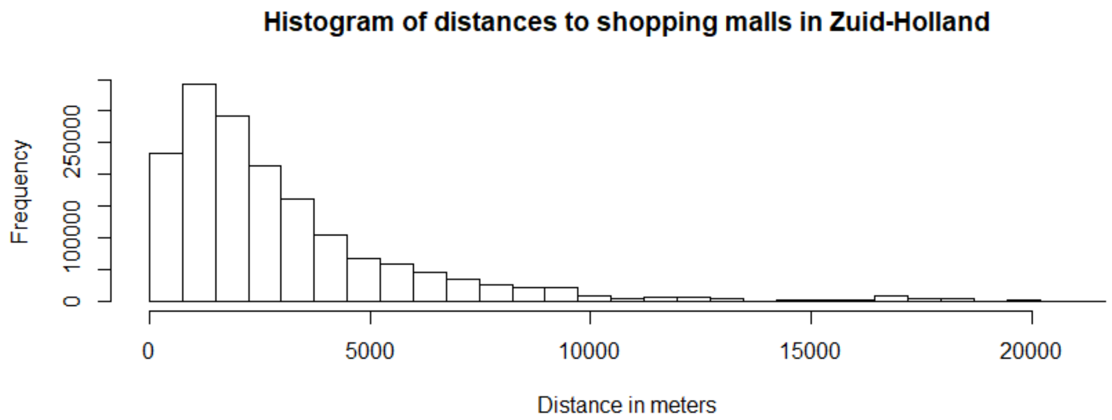
Interaction: log(living space) * factor district 'Wijk 14 Voordijkshoorn'	$7.61 \cdot 10^{-1}$	$2.32 \cdot 10^{-3}$
Interaction: log(living space) * factor district 'Wijk 16 Delftse Hout'	$7.81 \cdot 10^{-1}$	$9.15 \cdot 10^{-3}$
Interaction: log(living space) * factor district 'Wijk 22 Tanthof-West'	$7.39 \cdot 10^{-1}$	$2.41 \cdot 10^{-3}$
Interaction: log(living space) * factor district 'Wijk 23 Tanthof-Oost'	$7.46 \cdot 10^{-1}$	$2.45 \cdot 10^{-3}$
Interaction: log(living space) * factor district 'Wijk 24 Voorhof'	$7.51 \cdot 10^{-1}$	$2.22 \cdot 10^{-3}$
Interaction: log(living space) * factor district 'Wijk 25 Buitenhof'	$7.51 \cdot 10^{-1}$	$2.30 \cdot 10^{-3}$
Interaction: log(living space) * factor district 'Wijk 26 Abtswoude'	$8.44 \cdot 10^{-1}$	$1.66 \cdot 10^{-2}$
Interaction: log(living space) * factor district 'Wijk 27 Schieweg'	$7.61 \cdot 10^{-1}$	$3.01 \cdot 10^{-3}$
Interaction: log(living space) * factor district 'Wijk 28 Wippolder'	$7.49 \cdot 10^{-1}$	$2.32 \cdot 10^{-3}$
Interaction: log(living space) * factor district 'Wijk 29 Ruiven'	$7.13 \cdot 10^{-1}$	$4.92 \cdot 10^{-3}$

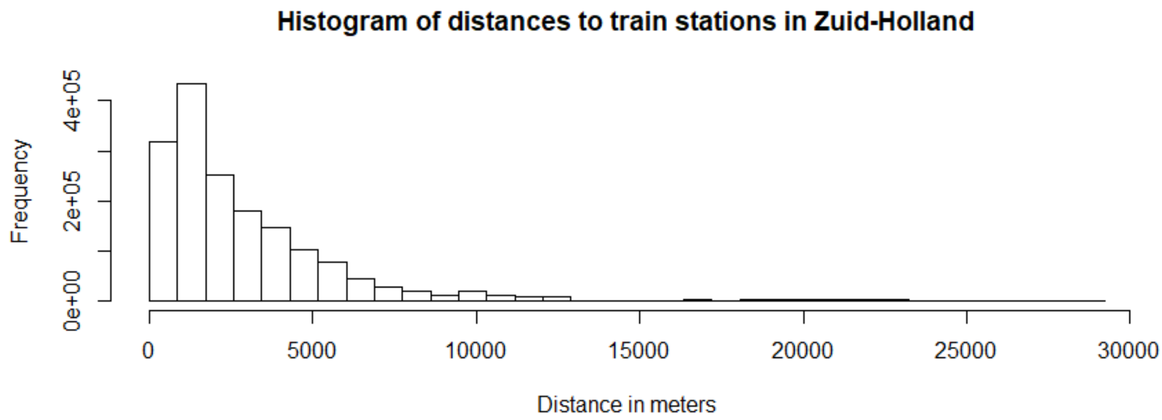
A.2. Provinces

A.2.1. Zuid-Holland

Histograms and groupings of distances to four facilities that have new categories compared to the model for Delft.







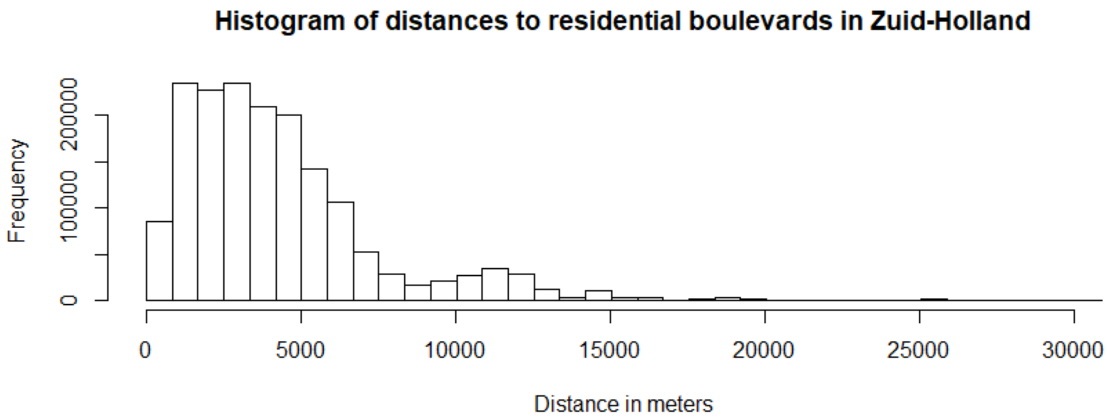


Table A.10: Grouping of the categories for supermarket distances in Zuid-Holland.

Categories for distance to nearest supermarket in Zuid-Holland	
Group name	Distance (x) range in meter
0	$x \leq 150$
1	$150 < x \leq 300$
2	$300 < x \leq 450$
3	$450 < x \leq 600$
4	$600 < x \leq 750$
5	$x > 750$

Table A.11: Grouping of the categories for shopping mall distances in Zuid-Holland.

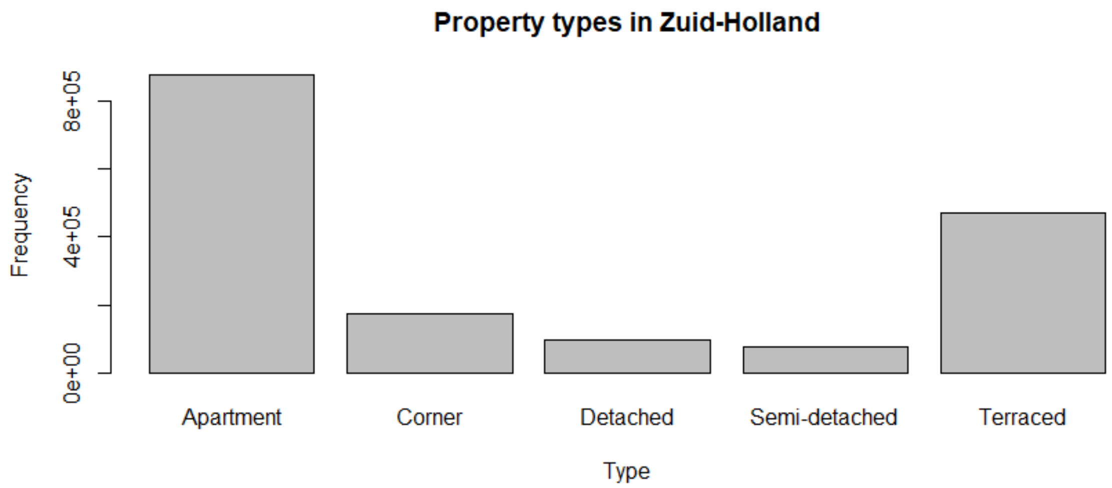
Categories for distance to nearest shopping mall in Zuid-Holland	
Group name	Distance (x) range in meter
0	$x \leq 750$
7.5	$750 < x \leq 1500$
15	$1500 < x \leq 2250$
22.5	$2250 < x \leq 3000$
30	$3000 < x \leq 3750$
37.5	$3750 < x \leq 4500$
45	$x > 4500$

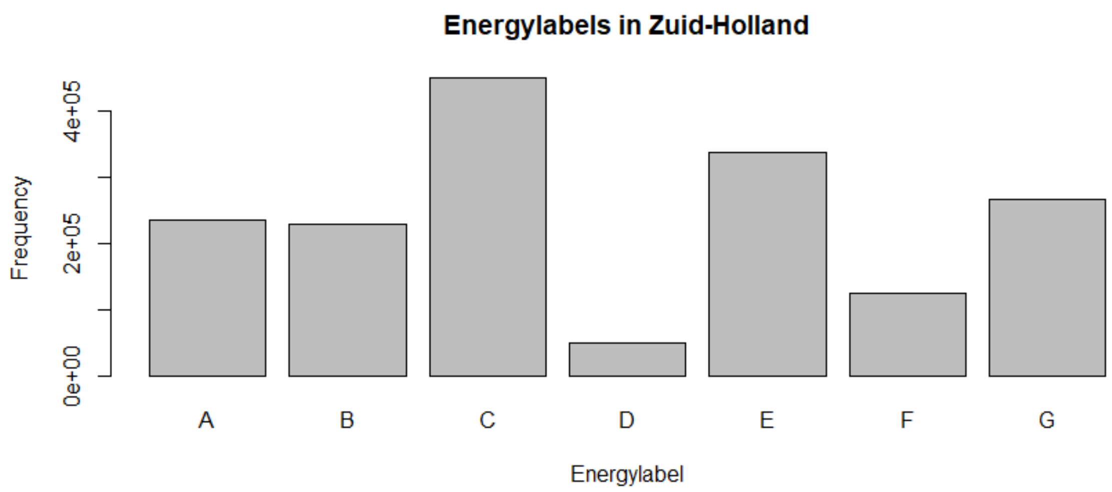
Table A.12: Grouping of the categories for train station distances in Zuid-Hollland.

Categories for distance to nearest train station in Zuid-Holland	
Group name	Distance (x) range in meter
0	$x \leq 750$
7.5	$750 < x \leq 1500$
15	$1500 < x \leq 2250$
22.5	$2250 < x \leq 3000$
30	$3000 < x \leq 3750$
37.5	$x > 3750$

Table A.13: Grouping of the categories for residential boulevard distances in Zuid-Holland.

Categories for distance to nearest residential boulevard in Zuid-Holland	
Group name	Distance (x) range in meter
0	$x \leq 1250$
12.5	$1250 < x \leq 2500$
25	$2500 < x \leq 3750$
37.5	$3750 < x \leq 5000$
50	$5000 < x \leq 6250$
62.5	$x > 6250$





A.3. Number of properties, WOZ values and asking prices

Here we give an overview of the number of properties for residential use we have for each group that we use in a model, the number of available WOZ values and the number of available asking prices.

Table A.14: Number of properties, WOZ values and asking prices

Group	Number of properties	Number of WOZ values	Number of asking prices
Zuid-Holland	1 695 303	1 556 670	87 602
Groningen, Friesland, Drenthe	795 658	624 254	45 339
Limburg	529 105	480 761	23 719
Noord-Brabant, Zeeland	1 322 298	1 042 337	63 195
Gelderland, Overijssel	1 408 570	1 273 571	73 317
Noord-Holland	1 371 744	1 213 017	60 315
Utrecht, Flevoland	749 147	647 528	43 149

List of Figures

List of Tables

Bibliography

[1] John Fox, *Applied Regression Analysis and Generalized Linear Models*, 2016, SAGE Publications, Third edition.

[2] Property types of cadastre available via:
`https://zakelijk.kadaster.nl/documents/20838/88047/Productbeschrijving+Woningtypering/a72e071a-e7af-4b93-aef2-a211de0f2056`

[3] C. Vuik and D.J.P. Lahaye, Lecture notes: *Scientific Computing (wi4201)*, 2019, TU Delft.
`http://ta.twi.tudelft.nl/nw/users/vuik/wi4201/wi4201_notes.pdf`

[4] Frank van der Meulen, Lecture notes: *Statistical Inference (wi4455)*, 2017, TU Delft.

[5] John Fox and Georges Monette, *Generalized Collinearity Diagnostics*, Journal of the American Statistical Association, 1992, Vol. 87, No. 417, pp. 178-183.

[6] E. de Jonge and M. van der Loo, *An introduction to data cleaning with R*, 2013, Statistics Netherlands.

[7] J. C. Gower, *A General Coefficient of Similarity and Some of Its Properties*, Biometrics, 1971, Vol. 27, No. 4, pp. 857-871.

[8] Alexander Kowarik and Matthias Templ, *Imputation with the R Package VIM*, 2016, Journal of Statistical Software, Volume 74, No. 7.

[9] (Determination of) WOZ values:
`www.wozwaardeloket.nl`

[10] Determination of WOZ values:
`https://www.waarderingskamer.nl/klopt-mijn-woz-waarde/totstandkoming-woz-waarde/`

[11] Lists of postal codes in the Netherlands. (Last checked on September 19, 2019.)

- `https://nl.wikipedia.org/wiki/Postcodes_in_Nederland`
- `https://nl.wikipedia.org/wiki/Lijst_van_postcodes_6000-6999_in_Nederland`
- `https://nl.wikipedia.org/wiki/Lijst_van_postcodes_5000-5999_in_Nederland#5800-5899`

[12] Average WOZ value for each municipality:
`https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37610/table?dl=10550`

[13] Jan de Haan, *Repeat Sales Methods*, Handbook on Residential Property Price Indices, 2013, OECD Publishing, Paris.

[14] Jan de Haan and Erwin Diewert, *Hedonic Regression Methods*, Handbook on Residential Property Price Indices, 2013, Eurostat, Luxembourg.

[15] Martin J. Bailey, Richard F. Muth and Hugh O. Nourse, *A Regression Method For Real Estate Price Index Construction*, 1963, Journal of the American Statistical Association, Vol. 58, No. 304, pp. 933-942.

[16] Karl E. Case and Robert J. Shiller, *Prices of Single Family Homes Since 1970: New Indexes for Four Cities*, 1987, NBER Working Paper, No. 2393.

[17] Karl E. Case and Robert J. Shiller, *The Efficiency of the Market for Single-Family Homes*, 1989, The American Economic Review, Vol. 79, No. 1, pp. 125-137.

[18] Chaitra Nagaraja, Lawrence Brown, Linda Zhao, *An autoregressive approach to house price modeling*, 2011, Ann. Appl. Stat., Volume 5, No. 1, pp. 124-149.

[19] Chaitra Nagaraja, Lawrence Brown, Susan Wachter, *House Price Index Methodology*, 2014, Journal of real estate literature, Vol 22, No. 1, pp. 1-21.

[20] Christian Hott and Pierre Monnin, *Fundamental Real Estate Prices: An Empirical Estimation with International Data*, National Centre of Competence in Research Financial Valuation and Risk Management, 2006, Working Paper No. 356.

[21] Qi Tu, Jan de Haan and Peter Boelhouwer, *House prices and long-term equilibrium in the regulated market of the Netherlands*, 2018, Housing Studies, Vol.33, No. 3, pp. 408-432.

[22] Nan Geng, *Fundamental Drivers of House Prices in Advanced Economies*, 2018, IMF Working Paper, No. 18/164.

[23] Information transaction prices via:
`https://www.kadaster.nl/web/kadaster.nl/producten/woning/koopsominformatie`

[24] Figure of the bulging rule from:
`https://freakonometrics.hypotheses.org/14967`

Data rescources

[25] BAG data via:
`https://nlextract.nl/downloads/`

[26] Energylabels and property types via:
`https://www.rvo.nl/onderwerpen/duurzaam-ondernemen/gebouwen/hulpmiddelen-tools-en-inspiratie-g`
`ep-online`

[27] Leefbaarometer data via:
`https://data.overheid.nl/data/dataset/leefbaarometer-2-0---meting-2016/resource/`
`c8a2b51b-ae3a-4fbd-909c-cfcf26db8fae`

[28] Neighborhood numbers (2018) in the Netherlands via:
`https://www.cbs.nl/nl-nl/maatwerk/2018/30/kerncijfers-wijken-en-buurten-2018`

[29] Neighborhood numbers (2017) in the Netherlands via:
`https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83765NED/table?ts=1565262427676`

[30] Regrouping municipalities by January 1st of 2019:
`https://www.rvig.nl/actueel/nieuws/2018/08/10/gemeentelijke-herindelingen-per-1-januari-2019`

[31] Regrouping municipalities by January 1st of 2018:
`https://www.rvig.nl/actueel/nieuws/2017/09/06/gemeentelijke-herindeling-per-1-januari-2018`

[32] 'Original data set' and asking prices via:
`https://www.jumba.nl`

[33] Open data Statistics Netherlands:
`https://opendata.cbs.nl/statline/#/CBS/nl/`

Software: R packages

Overview of the used R packages and functions for mathematical and statistical purposes:

Package	**Function(s)**
MASS	rlm(), studres(), polr()
MLmetrics	MAPE()
stats	lm(), alias()
L1pack	lad()
VIM	kNN()
caret	createDataPartition()
car	crPlots(), avPlots(), outlierTest(), vif()
robustbase	Mchi(), Mpsi(), Mwgt()

[34] Manuals of the R packages can be found via:
`https://cran.r-project.org/web/packages/available_packages_by_name.html`